SPECTRUM RECONSTRUCTION TECHNIQUE AND IMPROVED NAIVE BAYES MODELS FOR TEXT CLASSIFICATION PROBLEMS

A Dissertation
Presented to
The Academic Faculty

By

Zhibo Dai

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology

August 2020

SPECTRUM RECONSTRUCTION TECHNIQUE AND IMPROVED NAIVE BAYES MODELS FOR TEXT CLASSIFICATION PROBLEMS

Approved by:

Dr. Heinrich Matzinger, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Federico Bonetto
School of Mathematics
Georgia Institute of Technology

Dr. Ionel Popescu
School of Mathematics
Georgia Institute of Technology

Dr. Wenjing Liao
School of Mathematics
Georgia Institute of Technology

Dr. Tuo Zhao
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: April , 2020

Take the first step in faith. You don't have to see the whole staircase, just take the first step.

Ḿartin Luther King Jr.

For Kun, Eva and Lucky

ACKNOWLEDGEMENTS

This thesis could not have been successfully completed without the invaluable assistance of many individuals and institutes. The following list of acknowledgements is, by no means, exhaustive.

I would like to first express my appreciation and gratitude to my advisor, Dr Heinrich Matzinger, for his assistance, guidance and support. It is a great honor to work with him. I have learned a lot from him, no matter in research and in life. His inspiring guidance and creative ideas help me overcome many obstacles in my PhD research. Nothing can be achieved without his continuous support and encouragement during my PhD study.

Next, I would like to take the opportunity to say thank you to Dr. Ionel Popescu, Dr. Federico Bonetto, Dr. Wenjing Liao and Dr. Tuo Zhao for serving as my committee members. I would like to extend my sincere gratitude to Dr. Ionel Popescu for his numerous guidance and suggestion on my research. Also I would like to thanks Dr. Bonetto, Dr. Liao and Dr. Mayya Zhilova for their helpful suggestions on my oral exam, which inspires me to do more in my research.

I am also very much indebted to the School of Mathematics at Georgia Institute of Technology for providing me continuous financial support opportunities throughout my PhD studies. In particular, I would like to extend my gratitude to the English classes offered by Dr. Morag Burke, as well as the training and advice on teaching given by Ms. Klara Grodzinsky, these are all very valuable experience which I will never forget.

During my PhD career, I have made a lot of friends here, with whom I shared an unforgettable memory. For Jiaqi Yang, Xiao Liu and Hangfan Li, we entered the program in the same year and developed a strong friendship during these years. For Dr. Jiangning Chen, Juntao Duan, we have worked together for many years and I learned a lot from them on how to be a qualified PhD researcher. There are also many friends, He Zhang, Lingquan Ding, Liyan Nie, Qi Wang, who helped me a lot during my job hunting and internship.

I really appreciate all their help and friendship and would like to take this chance to say thank you to all.

I would also express my heartfelt thanks to my old friends. Dr. Yue Zhang, Dr. Cheng Chen and Guotao Chu motivate me a lot during my PhD. And I would like to extend my deep gratefulness to Xinrui Nie and Lujia Zhang for their kindly support and whenever I came across some difficult time, I can always seek help from them.

Last but not the least, my heartiest thanks flow to my family. My parents Min Lu, Wenting Dai, Cuimin Ji and Mengzhao Pan gave me the unconditional love and support. I would like to say thanks to my wife Kun Pan for her unconditional support and love and spending her life with me. Without her, I won't be able to achieve the success I have today. I love her so much. For my lovely daughter Eva, this thesis would be your first birthday gift.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

This thesis studies two topics. In the first part, we study the spectrum reconstruction technique. As is known to all, eigenvalues play an important role in many research fields and are foundation to many practical techniques such like PCA(Principal Component Analysis). We believe that related algorithms should perform better with more accurate spectrum estimation. There was an approximation formula proposed by[1], however, they didn't give any proof. In our research, we show why the formula works. And when both number of features and dimension of space go to ∞, we find the order of error for the approximation formula, which is related to a constant c-the ratio of dimension of space and number of features.

In the second part, we focus on some applications of Naive Bayes models in text classification problems. Especially we focus on two special situations: 1) there is insufficient data for model training; 2) partial label problem. We choose Naive Bayes as our base model and do some improvement on the model to achieve better performance in those two situations. To improve model performance and to utilize as many information as possible, we introduce a correlation factor, which somehow relax the conditional independence assumption of Naive Bayes. The new estimates are biased estimation compared to the traditional Naive Bayes estimate, but have much smaller variance, which give us a better prediction result.

CHAPTER 1

ANALYTICAL FORMULA FOR SPECTRUM RECONSTRUCTION

1.1 Introduction

Our research is about a simple analytical formula for the difference between the sample covariance and ground truth covariance spectrum of large multivariate normal data. We let both of the sample size and the dimension, in which the data lives, go to infinity at the same time. We show why a simple analytical approximation formula for the difference between sample spectrum and ground truth spectrum 1.4.25 and similarly 1.3.18 holds in certain cases. These formulas have already been introduced in [1], but are without a justification of why they should hold. In section 1.3, we show that the approximation 1.3.18 holds when a given condition 1.3.25 on the size of the sample holds. This condition is for eigenvalues, which are of somewhat larger order at least $O(n^{0.5})$ as we argue. Note that this result is not asymptotic and we use the result of Lounici and Koltchinskii [2] allowing to bound the error matrix in covariance estimation.

In Section 1.4, we consider the situation where the sample size is a large constant times the space dimension. We show that by taking the constant big enough, we get approximation 1.4.25 to hold as good as we want (relative error as small as we want) on the inside.

Let us first give the background of the problem: Assume that Z is a n by p data matrix. Assume, for example, that each column of Z is a point in a machine learning problem. Note that the product matrix Z^tZ contains all the inner products between columns. Here Z^t represents the transpose of Z. It is then easy to see that from Z^tZ we can find all the relative positions of the column vectors with respect to each other. We view them as vectors of $\mathbb{R}^n$. So, any machine learning algorithm, of which output depends only on the relative position of points to each other, would only need Z^tZ as input rather than Z given the input

points for the algorithm are the columns of Z.

We consider a random matrix Z with i.i.d. columns distributed each like a random vector $\vec{Z} = (Z_1, Z_2, \ldots, Z_p)$, which we assume to have zero expectation due to standardizing the data done before using most machine learning algorithms. Then, the *sample covariance matrix* is defined as:

$$\hat{\mathrm{COV}}[\vec{Z}] = \frac{Z^t \cdot Z}{n} \tag{1.1.1}$$

and is an unbiased estimate of the covariance $\mathrm{COV}[\vec{Z}]$. The covariance matrix

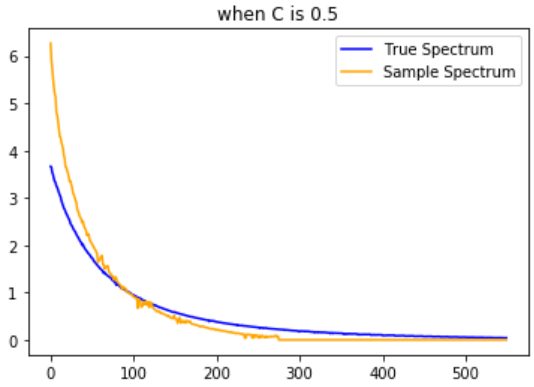
$$\mathrm{COV}[\vec{Z}] = (E[Z_i \cdot Z_j])_{ij} \tag{1.1.2}$$

will be called the *ground truth covariance*. Now, the sample covariance as estimate of the ground truth covariance is very bad as long as $n < p$ since then 1.1.1 is defective, which means it has some eigenvalues approximating 0 assuming $\mathrm{COV}[\vec{Z}]$ has no zero eigenvalues. This is the problem that affects many machine learning algorithms and it is also called the curse of dimensionality. In traditional statistics, one assumes p fixed while n goes to infinity. In modern high dimensional statistics, one lets $n = c \cdot p$, where c is a constant not depending on p. This implies that both n and p go to infinity at the same time, which is the situation we consider in our research. In the case that c is not large enough, as already mentioned the sample covariance is a very bad estimate of the true covariance since its small eigenvalues will be much smaller than the corresponding eigenvalues of ground truth covariance.

Now we will assume that we are dealing with multivariate normal vector $\vec{Z}$. We denote by σ_j^2, the j-th eigenvalue of the covariance $\mathrm{COV}[\vec{Z}]$ in descending order. So the spectrum of the covariance matrix is

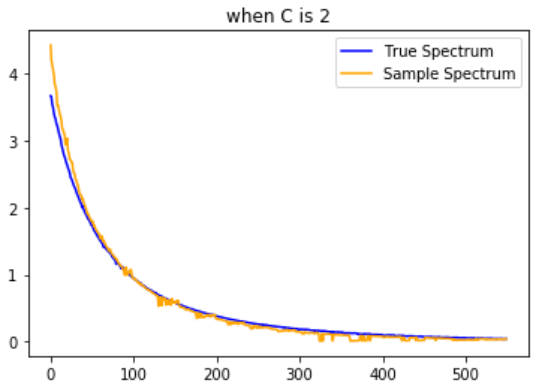
$$\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_p^2 \tag{1.1.3}$$

and the corresponding eigenvalues of the sample covariance will be denoted by $\hat{\sigma}_j^2$ and



(a)

Figure 1.1: Covariance Spectrum vs Ground Truth Spectrum when n = p/2



(a)

Figure 1.2: Covariance Spectrum vs Ground Truth Spectrum when n = 2*p

hence

$$\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2 \geq \ldots \geq \hat{\sigma}_p^2. \tag{1.1.4}$$

The topic of this chapter 1 is reconstruction of 1.1.3 when given only 1.1.4 in the context of normal data. Note that when both n and p go to infinity, the fluctuation of 1.1.4 is of smaller order than the values themselves. So for practical purposes we can consider the spectrum 1.1.4 to be non-random. In traditional statistics, where p is fixed as n goes to infinity the opposite is true. The unit-eigenvectors of the covariance matrix 1.1.2, are called *Principal Components*. When we represent the vector $\vec{Z}$ in the coordinate system of the Principal Components the new coordinates are uncorrelated. For a normal vector this implies independence. So, let

$$\vec{X} = (X_1, X_2, \ldots, X_p)$$

denote the vector $\vec{Z}$ expression in the Principal Components of $\mathrm{COV}[\vec{Z}]$. So the random vector $\vec{X}$ is a normal vector with independent normal components where

$$VAR[X_i] = \sigma_i^2$$

for $i = 1, 2, \ldots, p$. So, we will also express the data matrix in the coordinate system of the PCA, which means that each row of Z is going to be expressed in the basis of the PCA. Hence, we will not work with the data matrix Z, but instead with a data matrix X, where each row is an independent copy of $\vec{X}$. This means that X is a $n \times p$ matrix, with i.i.d rows where columns are also independent. In the j-th column we have normal entries with 0 expectation and variance σ_j^2. So, from here on the sample covariance is defined by

$$\hat{\mathrm{COV}}[\vec{X}] := \frac{X^t \cdot X}{n}$$

and the ground truth covariance is the diagonal matrix

$$\mathrm{COV}[\vec{X}] = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \dots & & & & \\ 0 & 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

We will also designate the ground truth covariance by Σ_p and hence

$$\Sigma_p := \mathrm{COV}[\vec{X}]$$

Now let $\Sigma_p^{1/2}$ be the square root of Σ_p::

$$\Sigma_p^{1/2} = \sqrt{\mathrm{COV}[\vec{X}]} = \begin{pmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \dots & & & & \\ 0 & 0 & 0 & \dots & \sigma_p \end{pmatrix}$$

Then the data matrix X has same distribution as $N \cdot \Sigma_p^{1/2}$, where N is a n times p matrix of i.i.d. standard normal entries. Hence, we can write the sample covariance as

$$\hat{\mathrm{COV}}[\vec{X}] = \frac{X^t \cdot X}{n} = \frac{\Sigma_p^{1/2} N^t \cdot N \Sigma_p^{1/2}}{n} \tag{1.1.5}$$

Now, we use the property that the product of square matrices $A \cdot B$ has same spectrum as $B \cdot A$. We let A be the matrix$\Sigma_p^{1/2}$, and B be $N^t \cdot N\Sigma_p^{1/2}$. So, after applying the rule that

AB and BA have same spectrum we find that the spectrum of

$$BA = \frac{(N^t \cdot N)}{n} \cdot \Sigma_p \tag{1.1.6}$$

is identical with the spectrum of the sample covariance 1.1.5, which is AB.

Now free probability theory tells us that the spectrum of the product of two independent symmetric random square matrices converges as long as each matrix's spectrum converges. Actually in order to have the convergence, we need random matrices to be unitary invariant, which is indeed the case here as $N^t N$ satisfies the condition. For this we let the dimension of the matrices go to infinity. During that convergence process, their spectrum converges to a "finite" limiting distribution. The limit of the product's spectrum is called the *Free Product* of the limiting distributions of each spectrum taken separately. We can apply this to the product on the right side of 1.1.6 as long as the spectrum of Σ_p converges, when p goes to infinity. In that case, the product must converge to a free product. That is the free product of the limiting distribution for Σ_p, with that of $\frac{(N^t \cdot N)}{n}$. In 1967, Vladimir Marčenko and Leonid Pastur [3] successfully constructed the limiting law of $\frac{(N^t \cdot N)}{n}$, which is now named after the authors, namely Marchenko–Pastur distribution. This is the case when both n and p go to infinity at the same time. Therefore, $\frac{n}{p}$ converges to a non-zero fixed limit, which we denote by c. The limiting law depends on c.

So, from our explanation of free probability, in the case of $\mathrm{COV}[X]$'s spectrum admitting a limiting law F^{Σ}, the sample spectrum is a free product of Marčenko-Pastur law and F^{Σ}. By computing the S-transform explicitly, one can obtain a formula of the limiting law of the sample spectrum. See Bai and Yin [4], Yin, Bai and Krishnaiah [5], Silverstein [6], and many others. The main result is summarized as follows:

Theorem 1.1.1. *Given the following conditions,*

1. *Suppose entries of $N_p = (N_{i,j})_{n \times p}$ are i.i.d. real random variables for all p.*

2. $E[N_{1,1}] = 0$, $E[|N_{1,1}|^2] = 1$.

3. *Let* $n/p \to c > 0$ *as* $p \to \infty$.

4. *Let* Σ_p $(p \times p)$ *be non-negative definite symmetric random matrix with spectrum distribution* F^{Σ_p} *(If* $\{\lambda_i\}_{1 \le i \le p}$ *are the eigenvalues of* Σ_p*, then* $F^{\Sigma_p} = \sum_1^p \frac{1}{p}\delta_{\lambda_i}(x)$*) such that* F^{Σ_p} *almost surely converges weakly to* F^{Σ} *on* $[0, \infty)$.

5. N_p *and* Σ_p *are independent.*

then the spectrum distribution of $W_p = \frac{1}{n}\Sigma_p^{1/2} N_p^T N_p \Sigma_p^{1/2}$*, denoted as* F^{W_p} *almost surely converges weakly to* F^W*.* F^W *is the unique probability measure whose Stieltjes transform* $m(z) = \int \frac{dF^W(x)}{x-z}$*,* $z \in \mathbb{C}^+$ *satisfies the equation*

$$-\frac{1}{m} = z - c\int \frac{t}{1+tm} dF^{\Sigma}(t) \quad \forall z \in \mathbb{C}^+ \tag{1.1.7}$$

So, recall the topic of this chapter is reconstructing the ground truth spectrum given only the spectrum of the sample covariance. Currently most methods for this problem are "free-probability based". That is they attempt to solve the equation 1.1.7 to get an estimator of the true spectrum F^{Σ}. Take the sample covariance spectrum as if it would be the limit, which is taking F^{W_p} for the distribution F^W in order to get the Stieltjes transform $m(z)$. So, instead of $m(z)$, we use $m_p(z)$ in 1.1.7. Then solve, which is finding F^{Σ} solving 1.1.7 and pretending it is F^{Σ_p}. In such an approach one hopes that there is only a little difference between the observed spectral distribution for a given p and the limiting distributions. Such an approach based on free probability theory was pioneered by EI Karoui [7], and then Bai etc. [8], and recently by Ledoit and Wolf [9]and [10]. It's not surprising that as dimensions grow, consistency is achieved by the free probability approach. But a disadvantage is that the recovered spectrum is still far from the true spectrum for small or moderate size of p since the method operates as if the data given would already be in the "free-probability limit". Another problem with the free probability approach is that the spectrum of the ground truth covariance needs to converge for free probability to be applicable. However,

in real data, there is always different order eigenvalues: some have order $O(1)$ and some have order $O(p)$.

In[1], Matzinger etc. proposes 2 methods to reconstruct population spectrum based on sample spectrum. The first one is a simple algebraic formula to reconstruct population eigenvalues given sample eigenvalues. It is our formula 1.4.25. Unless the structure of population spectrum is too flat, this estimation performs well. Note that our formula can be interpreted as: the relative error between sample spectrum and ground truth spectrum is approximately the Stieltjes transform of sample spectrum. Indeed, our approximation 1.4.25 can be rewritten:

$$c \cdot \frac{\hat{\sigma}_i^2 - \sigma_i^2}{\sigma_i^2} \approx -\frac{1}{p} \sum_{j \neq i} \frac{\hat{\sigma}_j^2}{\hat{\sigma}_j^2 - \hat{\sigma}_i^2} = -1 - \frac{\hat{\sigma}_i^2}{p} \sum_{j \neq i} \frac{1}{\hat{\sigma}_j^2 - \hat{\sigma}_i^2} = -1 - \hat{\sigma}_i^2 \int \frac{1}{x - \hat{\sigma}_i^2} dF^{W_p}(x), \tag{1.1.8}$$

where for the integral over the spectral measure F^{W_p} we make the convention that we leave out the atom at $\hat{\sigma}_i^2$ since otherwise we would have 0 in the denominator of the summation. (Also, as usual we have $n = c \cdot p$). Note that on the very right side of 1.1.8, we have the Stieltjes transform of the empirical distribution of the sample covariance spectrum. The second method proposed by Matzinger etc.[1] is a fixed point method. The second approach is more computation-expensive but achieves a more accurate estimate, and we won't treat it here. Our research focus on the first approach proposed by Matzinger etc all [1]. We do a deeper analysis on the error term and theoretically show it is negligible under certain condition. We also present a similar formula 1.3.18 but where sample covariance and covariance spectrum are inverted.

In most situations, researchers use the spectrum extracted from a sample matrix, especially the sample covariance matrix, which brings in some error due to sample estimation bias. Therefore, estimating the eigenvalues of a population covariance matrix from a sample covariance matrix is of fundamental importance. The population spectrum will provide us more accurate essential information about the structure of the data problem[11].

1.2 Related Work

Both of eigenvalues and eigenvectors have significant influence in mathematics and real life. Theoretically, they can be applied in linear algebra, differential operators and dynamic equations such as matrix diagonalization, eigen decomposition, eigenvector-eigenvalue identity and solving differential equations. Apart from mathematics, researchers also utilize the properties of eigenvalue and eigenvector in Schrödinger equation[12], geology[13] and vibration analysis. The widely known application would be Principal Component Analysis(PCA)[14], which is used in dimension reduction[15], feature selection[16, 17, 18], K-means clustering[19] and general text classification problems[20, 21, 22, 23].

In most situations, researchers use the spectrum extracted from a sample matrix, especially the sample covariance matrix, which brings in some error due to sample estimation bias. Therefore, estimating the eigenvalues of a population covariance matrix from a sample covariance matrix is of fundamental importance. The population spectrum will provide us more accurate essential information about the structure of the data. There are a family of researches on spectrum reconstruction algorithms, which attempt to discretize and adapt the free probability infinite dimensional recovery. The idea was first introduced by [7] based on the Marčenko–Pastur equation. In[5], authors show that the spectral distribution of a central multivariate matrix converges to a a limit distribution in probability. Then in[4] the convergence of the spectral distribution of the sample covariance matrix to the semicircle law was proved given the assumption that $X_p = [X_{ij}]_{p*n}$ has iid entries and $E(X_{11}^4) < \infty, var(X_{11}) = 1$. A similar result on strong convergence of the empirical distribution of eigenvalues was proved in 1995 by [6]. Recently, in[9][10] authors propose a novel estimate of the population eigenvalues which is consistent under large-dimensional asymptotics regardless of whether or not they are clustered, and that also performs well in

finite sample. They find the estimate by solving the following optimization problem:

$$\hat{(}\tau_n) = \underset{t\in[0,\infty)^p}{\arg\min} \frac{1}{p}\sum_{i=1}^{p}[q_{n,p}^i(t) - \lambda_{n,i}]^2$$

where $Q_{n,p}(t) = (q_{n,p}^1(t), \ldots, q_{n,p}^p(t))^t$ is the nonrandom QuEST function. And this convergence is almost surely convergence. In[24] researchers show another new method founded on a meaningful generalization of the seminal Marcenko-Pastur equation, originally defined in the complex plan, to the real line.

Shrinkage is also one of methods to reconstruct population spectrum. The idea was pioneered by Stein[25]. See also Bickel[26] and Donoho[27]. Another type of approach is based on the moments of the spectral distributions[28], which shows a theoretically optimal and computationally efficient algorithm for recovering the moments of the population eigenvalues. Finally, there are also Physicists Burda, Gorlich and Jarosz, working on this problem[11].

1.3 Case of Larger Order Eigenvalues

In this subsection we show the approximation formula 1.3.18 to hold when the condition 1.3.25 is satisfied. So we first write down a three dimensional vector but the formula will still be useful in high dimension case. Now we have a sequence of i.i.d. vectors with 0 expectation:

$$\vec{X}, \vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$$

where $\vec{X}_i = (X_i, Y_i, Z_i)$ and $\vec{X} = (X, Y, Z)$

We will assume that

$$\begin{aligned} & E[\vec{X}] \\ &= E[(X,Y,Z)] \\ &= E[\vec{X}_i] \\ &= (E[X_i], E[Y_i], E[Z_i]) \\ &= (0,0,0) \end{aligned}$$

We will explain later why in many applications, this assumption is realistic. We assume that X_i, Y_i and Z_i are independent of each other. Hence, the covariance matrix is given by

$$\mathrm{COV}[\vec{X}] = \begin{bmatrix} \sigma_X^2 & 0 & 0 \\ 0 & \sigma_Y^2 & 0 \\ 0 & 0 & \sigma_Z^2 \end{bmatrix} = \begin{bmatrix} E[X^2] & E[XY] & E[XZ] \\ E[YX] & E[Y^2] & E[YZ] \\ E[ZX] & E[ZY] & E[Z^2] \end{bmatrix}$$

Now recall the Central Limit Theorem: Assume we have variables $W_1, W_2, \ldots$, which are i.i.d, then we have for n large enough, the properly re-scaled sum is approximately standard normal:

$$\frac{W_1 + W_2 + \ldots + W_n - nE[W_1]}{\sqrt{n}\sigma} \approx \mathcal{N}(0,1)$$

the goal is to figure out how precise our estimates for the eigenvalues and eigenvectors are. Since the expectation is 0, in our estimate of the covariance matrix we can leave the part which estimates the expectation out. Then we use the following estimate for the covariance matrix:

$$\hat{\mathrm{COV}}[\vec{X}] = \begin{bmatrix} \frac{X_1^2+\ldots+X_n^2}{n} & \frac{X_1Y_1+\ldots+X_nY_n}{n} & \frac{X_1Z_1+\ldots+X_nZ_n}{n} \\ \frac{Y_1X_1+\ldots+Y_nX_n}{n} & \frac{Y_1^2+\ldots+Y_n^2}{n} & \frac{Y_1Z_1+\ldots+Y_nZ_n}{n} \\ \frac{Z_1X_1+\ldots+Z_nX_n}{n} & \frac{Z_1Y_1+\ldots+Z_nY_n}{n} & \frac{Z_1^2+\ldots+Z_n^2}{n} \end{bmatrix}$$

We can now apply the Central Limit Theorem to all entries of the estimated covariance

matrix above. For example let's take W_i to be $W_i = X_i Y_i$. Then

$$\begin{aligned}\frac{X_1Y_1 + \ldots + X_nY_n}{n} - E[X_1Y_1] &= \frac{1}{\sqrt{n}} \frac{W_1 + W_2 + \ldots + W_n - E[W_1]}{\sqrt{n}} \\ &\approx \sigma_{W_1} \frac{\mathcal{N}(0,1)}{\sqrt{n}} \\ &= \frac{\sigma_{X_1}\sigma_{Y_1}}{\sqrt{n}} \mathcal{N}(0,1) \end{aligned} \tag{1.3.1}$$

So take the difference E between the estimated covariance matrix and the real one, being called the covariance estimation matrix:

$$\begin{aligned} E &= \hat{\mathrm{COV}}[\vec{X}] - \mathrm{COV}[\vec{X}] \\ &= \begin{bmatrix} \frac{X_1^2+\ldots+X_n^2}{n} - E[X^2] & \frac{X_1Y_1+\ldots+X_nY_n}{n} - E[XY] & \frac{X_1Z_1+\ldots+X_nZ_n}{n} - E[XZ] \\ \frac{Y_1X_1+\ldots+Y_nX_n}{n} - E[YX] & \frac{Y_1^2+\ldots+Y_n^2}{n} - E[Y^2] & \frac{Y_1Z_1+\ldots+Y_nZ_n}{n} - E[YZ] \\ \frac{Z_1X_1+\ldots+Z_nX_n}{n} - E[ZX] & \frac{Z_1Y_1+\ldots+Z_nY_n}{n} - E[ZY] & \frac{Z_1^2+\ldots+Z_n^2}{n} - E[Z^2] \end{bmatrix} \end{aligned}$$

With the Central Limit Theorem applied to each of the entries of the last matrix above in the same way as in 1.3.1.Now, let N_{ij} be the re-scaled i, j-th entry of our covariance estimation error matrix. Hence,

$$N_{12} = \frac{\sqrt{n}E_{12}}{\sigma_X\sigma_Y}, N_{13} = \frac{\sqrt{n}E_{13}}{\sigma_X\sigma_Z}, N_{23} = \frac{\sqrt{n}E_{23}}{\sigma_Y\sigma_Z}$$

and

$$N_{11} = \frac{\sqrt{n}E_{11}}{\sigma_X^2}, N_{22} = \frac{\sqrt{n}E_{22}}{\sigma_Y^2}, N_{33} = \frac{\sqrt{n}E_{33}}{\sigma_Z^2},$$

whilst $N_{ij} = N_{ji}$. By definition, the term N_{ij} has expectation 0 and standard deviation 1.

Clearly as n goes to ∞, the N_{ij} is asymptotically standard normal. With this notation:

$$\hat{\mathrm{COV}}[\vec{X}] - \mathrm{COV}[\vec{X}] = \frac{1}{\sqrt{n}} \begin{bmatrix} \sigma_X^2 N_{11} & \sigma_X \sigma_Y N_{12} & \sigma_X \sigma_Z N_{13} \\ \sigma_Y \sigma_X N_{21} & \sigma_Y^2 N_{22} & \sigma_Y \sigma_Z N_{23} \\ \sigma_Z \sigma_X N_{31} & \sigma_Z \sigma_Y N_{32} & \sigma_Z^2 N_{33} \end{bmatrix} \tag{1.3.2}$$

Also, note that the terms $N_{11}, N_{22}, N_{33}, N_{12}, N_{13}, N_{23}$ are all pairwise uncorrelated. For example:

$$\mathrm{COV}(XY, XZ) = E[XYXZ] - E[XY] \cdot E[XZ] = E[X^2]E[Y]E[Z] - E[X]E[Y]E[X]E[Z] = 0$$

Hence,

$$\begin{aligned} &\mathrm{COV}\left(\frac{X_1Y_1 + \ldots + X_nY_n}{n}, \frac{X_1Z_1 + \ldots + X_nZ_n}{n}\right) \\ &= \frac{1}{n^2} \sum_{i,j} \mathrm{COV}(X_iY_i, X_jZ_j) = \frac{1}{n^2} \sum_i \mathrm{COV}(X_iY_i, X_iZ_i) \\ &= 0 \end{aligned}$$

Next we are going to establish the formula for the estimated eigenvalue and eigenvectors of the covariance matrix. Again, the estimated eigenvalues and eigenvectors are simply the eigenvectors and eigenvalues of the estimated covariance matrix. We assume σ_X, σ_Y and σ_z all have different values. Let A denote the covariance matrix, E again the error-matrix, which is the difference between the estimated and the true covariance matrix. Let $\vec{\mu} = (1, 0, 0)^T$ be the first eigenvector of $A = \mathrm{COV}[\vec{X}]$. Let $\lambda = \sigma_X^2$ denote the first eigenvalue of the covariance matrix A and let $\lambda + \Delta\lambda$ denote the first eigenvalue of the estimated covariance matrix.

So the estimated covariance matrix is $A + E$, hence the true covariance matrix plus a "perturbation" E. Let $\vec{v} = \vec{\mu} + \Delta\vec{\mu}$ be the first eigenvector for the estimated covariance matrix and assume that $\Delta\vec{\mu}$ is orthogonal to μ. Hence $\Delta\vec{\mu} = (0, \Delta\mu_Y, \Delta\mu_Z)^T$. With these

notations, we have:

$$(A+E)(\vec{\mu}+\Delta\vec{\mu}) = (\lambda+\Delta\lambda)(\vec{\mu}+\Delta\vec{\mu}). \tag{1.3.3}$$

Also, since $\vec{\mu}$ is an eigenvector of A, we have:

$$A\vec{\mu} = \lambda\vec{\mu} \tag{1.3.4}$$

Subtracting equation 1.3.3 from 1.3.4, we find:

$$(A - I\lambda)\Delta\vec{\mu} = -E\vec{\mu} + \Delta\lambda\vec{\mu} + -E\Delta\vec{\mu} + \Delta\lambda\Delta\vec{\mu}. \tag{1.3.5}$$

we find the following exact equation:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_Y^2 - \sigma_X^2 - \Delta\lambda & 0 \\ 0 & 0 & \sigma_Z^2 - \sigma_X^2 - \Delta\lambda \end{bmatrix} \begin{bmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$
$$= \frac{-1}{\sqrt{n}} \begin{bmatrix} \sigma_X^2 N_{11} \\ \sigma_X\sigma_Y N_{21} \\ \sigma_X\sigma_Z N_{31} \end{bmatrix} + \begin{bmatrix} \Delta\lambda \\ 0 \\ 0 \end{bmatrix}$$
$$- \frac{1}{\sqrt{n}} \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_Y^2 N_{22} & \sigma_Y\sigma_Z N_{23} \\ 0 & \sigma_Y\sigma_Z N_{32} & \sigma_Z^2 N_{33} \end{bmatrix} \begin{bmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$
$$- \frac{1}{\sqrt{n}} \begin{bmatrix} 0 & \sigma_X\sigma_Y N_{12} & \sigma_X\sigma_Z N_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$

the above equation for matrices can be separated into two parts. First the single equation

for $\Delta\lambda$:

$$\Delta\lambda = \frac{1}{\sqrt{n}}\sigma_X^2 N_{11} + \frac{\sigma_X}{\sqrt{n}}(\sigma_Y N_{12}\Delta\mu_Y + \sigma_Z N_{13}\Delta\mu_Z), \tag{1.3.6}$$

which we will use to determine $\Delta\lambda$. Then the $p-1$ dimensional equation for $\Delta\vec{\mu}$ given as follows:

$$\begin{bmatrix} \sigma_Y^2 - \sigma_X^2 - \Delta\lambda & 0 \\ 0 & \sigma_Z^2 - \sigma_X^2 - \Delta\lambda \end{bmatrix} \begin{bmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$
$$= \frac{-1}{\sqrt{n}} \begin{bmatrix} \sigma_X\sigma_Y N_{21} \\ \sigma_X\sigma_Z N_{31} \end{bmatrix} - \frac{1}{\sqrt{n}} \begin{bmatrix} \sigma_{Y^2} N_{22} & \sigma_Y\sigma_Z N_{23} \\ \sigma_Y\sigma_Z N_{32} & \sigma_{Z^2} N_{33} \end{bmatrix} \begin{bmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$

If $\Delta\lambda$ is given, we can solve the above equation for $\Delta\vec{\mu} = (\Delta\mu_Y, \Delta\mu_Z)$ to find:

$$\Delta\vec{\mu} = \begin{bmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix}$$
$$= \left(I - \frac{-1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} & 0 \\ 0 & \frac{1}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{Y^2} N_{22} & \sigma_Y\sigma_Z N_{23} \\ \sigma_Y\sigma_Z N_{32} & \sigma_{Z^2} N_{33} \end{bmatrix} \right)^{-1}$$
$$\cdot \frac{-1}{\sqrt{n}} \begin{bmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} N_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda} N_{31} \end{bmatrix}$$

where I is the identity matrix. Now, let D_1 be the matrix

$$D_1 := - \begin{bmatrix} \frac{1}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} & 0 \\ 0 & \frac{1}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda} \end{bmatrix} \tag{1.3.7}$$

and let E_1 be the matrix obtained from E by deleting the first column and the first row:

$$E_1 = \begin{bmatrix} E_{22} & E_{23} \\ E_{32} & E_{33} \end{bmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} \sigma_Y^2 N_{22} & \sigma_Y\sigma_Z N_{23} \\ \sigma_Y\sigma_Z N_{32} & \sigma_Z^2 N_{33} \end{bmatrix} \tag{1.3.8}$$

Therefore, we have

$$\Delta\vec{\mu} = \begin{bmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix} = -\frac{1}{\sqrt{n}}(I - D_1E_1)^{-1} \begin{bmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2-\sigma_X^2-\Delta\lambda}N_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2-\sigma_X^2-\Delta\lambda}N_{31} \end{bmatrix} \tag{1.3.9}$$

Now when the spectral norm of D_1E_1 is less than 1, then we get the formula:

$$(I - D_1E_1)^{-1} = I + D_1E_1 + (D_1E_1)^2 + (D_1E_1)^3 + \dots$$

In the case where D_1E_1 has spectral norm quite a bit less than 1, we can approximate $(I - D_1E_1)^{-1}$ by I and find that:

$$\Delta\vec{\mu} = \begin{bmatrix} \Delta\mu_Y \\ \Delta\mu_Z \end{bmatrix} \approx -\frac{1}{\sqrt{n}} \begin{bmatrix} \frac{\sigma_X\sigma_Y}{\sigma_Y^2-\sigma_X^2-\Delta\lambda}\hat{N}_{21} \\ \frac{\sigma_X\sigma_Z}{\sigma_Z^2-\sigma_X^2-\Delta\lambda}\hat{N}_{31} \end{bmatrix}$$

with the relative error in that approximation being less than $\frac{|D_1E_1|}{1-|D_1E_1|}$. We can now plug the above formula into 1.3.6 and get:

$$\Delta\lambda \approx \frac{1}{\sqrt{n}}\sigma_X^2N_{11} - \frac{\sigma_X^2}{\sqrt{n}}(\frac{\sigma_Y^2N_{12}^2}{\sigma_Y^2 - \sigma_X^2 - \Delta\lambda} + \frac{\sigma_Z^2N_{13}^2}{\sigma_Z^2 - \sigma_X^2 - \Delta\lambda}), \tag{1.3.10}$$

If we don't take a three dimensional vector $\vec{X} = (X, Y, Z)$ but instead a p-dimensional $\vec{X} = (X_1, X_2, \dots, X_p)$ with independent normal entries, we find that the formula 1.3.10 becomes

$$\Delta\lambda = \frac{1}{\sqrt{n}}\sigma_{X_1}^2N_{11} - \frac{\sigma_{X_1}^2}{\sqrt{n}}\left(\frac{\sigma_{X_2}^2N_{12}^2}{\sigma_{X_2}^2 - \sigma_{X_1}^2 - \Delta\lambda} + \frac{\sigma_{X_3}^2N_{13}^2}{\sigma_{X_3}^2 - \sigma_{X_1}^2 - \Delta\lambda} + \dots + \frac{\sigma_{X_p}^2N_{1p}^2}{\sigma_{X_p}^2 - \sigma_{X_1}^2 - \Delta\lambda}\right) \tag{1.3.11}$$

where

$$N_{1j} = \sqrt{n}\frac{E_{1j}}{\sigma_{X_1}\sigma_{X_j}}$$

Also, here $\sigma_{X_1}^2 + \Delta\lambda$ represents the eigenvalue of the sample covariance, to which we

compare the first eigenvalue σ_1^2 of the true covariance. Note that we did not use the order of eigenvalues, hence σ_1^2 could be any eigenvalue of the true covariance. Also, for our formula 1.3.11 we don't need the fact that the ground truth eigenvalues are ordered to be held. So, to simplify notation let us denote by σ_j^2 the variance $VAR[X_j] = \sigma_{X_j}^2$ which is also the j-the eigenvalue of the covariance matrix $COV[\vec{X}]$. Now, $\sigma_{X_1}^2 + \Delta\lambda$ can denote any eigenvalue of the sample covariance matrix. So, let us write

$$\hat{\sigma}_1^2 > \hat{\sigma}_2^2 > \ldots > \hat{\sigma}_p^2$$

for the eigenvalues of the sample covariance. Let i^* be the index of the sample covariance eigenvalue to which we compare the eigenvalue σ_i^2. Hence, with this generalisation 1.3.11 becomes:

$$\Delta\lambda_i = \sigma_{i^*}^2 - \sigma_i^2 \tag{1.3.12}$$

and 1.3.11 can be rewritten as

$$\Delta\lambda_i \approx -\frac{\sigma_i^2}{\sqrt{n}} \sum_{j \neq i} \frac{\sigma_j^2 N_{ij}^2}{\sigma_j^2 - \hat{\sigma}_{i^*}^2}, \tag{1.3.13}$$

where we left out the term $\frac{1}{\sqrt{n}}\sigma_{X_1}^2 N_{11}$ which is a smaller order term. Now, for the above 1.3.13 to be useful, we need

$$\sigma_j^2 - \hat{\sigma}_{i^*}^2 \tag{1.3.14}$$

not to be too small. Indeed if for example the sample eigenvalue $\hat{\sigma}_{i^*}^2$ is equal to one of the ground truth eigenvalues σ_j^2 with $j \neq i$, then we would have zero in the denominator of one of the terms in the sum in 1.3.13. The only way to control this is to take i^* to be the index of the sample eigenvalue which comes closest to σ_i^2. In this way, we guarantee that in our sum 1.3.13, the expression 1.3.14 does not get beneath 0.5 times the spectral gap number i

of the ground truth. The spectral gap is defined as follows:

$$\texttt{spectral\_gap}_i := min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\} \tag{1.3.15}$$

and so we take i^* to be defined as:

$$i^* := \arg\min_j (j \mapsto |\sigma_i^2 - \hat{\sigma}_j^2|). \tag{1.3.16}$$

The other reason for taking such i^* is that the matrix D_1 defined before may also explode due to 1.3.14 being uncontrolled small. Now, defining $\Delta\lambda_i$ using i^* leads t a meaningless formula in the $O(1)$ part of the spectrum, which has $O(p)$ eigenvalues. So their distances should be about $O(1/p)$. Then, if we choose to compare σ_i^2 with the sample eigenvalues that come closest, we get $\Delta\lambda_i$ to be meaningless: the difference of sample spectrum and ground truth is $O(1)$ in that area. However, since the eigenvalues build a continuum, their distances are infinitesimal, which means we get $\Delta\lambda_i$ defined in 1.3.6 to be infinitesimal and not $O(1)$. So, the current section is for an error of the spectrum larger than $O(p^{0.5})$. (Compare with the remark at the very end of this section). The case of the eigenvalues of order $O(1)$ is treated in the next section.

Now, we assume that the ground truth eigenvalues are spaced very regularly. So that

$$\sigma_i^2 - \sigma_{i+1}^2, \sigma_i^2 - \sigma_{i+2}^2, \sigma_i^2 - \sigma_{i+3}^2, \ldots$$

behave like the sequence $\Delta_i, 2 \cdot \Delta_i, 3 \cdot \Delta_i, \ldots$, where $\Delta_i > 0$ is the spectral gap defined in 1.3.15. We assume the same thing to be held on the left side of i. Note that we have the series

$$\frac{1}{\Delta_i} + \frac{1}{2\Delta_i} + \frac{1}{3\Delta_i} + \ldots = \infty$$

is divergent. This implies that terms

$$\frac{\sigma_i^2}{\sigma_i^2 - \hat{\sigma}_{i^*}^2}, \frac{\sigma_{i+1}^2}{\sigma_{i+1}^2 - \hat{\sigma}_{i^*}^2}, \frac{\sigma_{i+2}^2}{\sigma_{i+2}^2 - \hat{\sigma}_{i^*}^2}, \dots$$

behave like the terms of a divergent series. Hence none of its terms dominates the sum

$$\sum_{j \geq i} \frac{\sigma_i^2}{\sigma_j^2 - \hat{\sigma}_{i^*}^2}$$

This has a practical importance for the expression on the right of approximation 1.3.13: the expected value dominates the fluctuation. This means that we can replace the standard normal random variables square $\mathcal{N}_{ij}^2$ by their expectation 1 and this causes only a smaller order change. The reason is as follows: consider a sum

$$\sum_j a_j N_j^2 \tag{1.3.17}$$

where N_j^2's are independent standard normals squared and the a_j's are constants. Then the expectation of 1.3.17 is $\sum_j a_j$ and dominates the sums standard deviation as soon as the sum $\sum_j a_j$ dominates any of its terms a_j Take now a_j to be $\frac{\sigma_j^2}{\sigma_j^2 - \hat{\sigma}_{i^*}^2}$. The condition that none of the a_j dominates the sum is satisfied due to the series being divergent, We act also as if $\hat{\sigma}_{i^*}^2$ would not be random. Hence, in the sum 1.3.13 we can replace the standard normal square $N_{I_j}^2$ by their expected value 1 and this will only cause a smaller order change. Hence, given the condition1.3.25, we finally obtain our result by replacing in 1.3.13 $\mathcal{N}_{ij}^2$ by 1 to get the approximation formula:

$$\boxed{\Delta\lambda_i \approx -\frac{\sigma_i^2}{n} \cdot \sum_{j \neq i}^{p} \frac{\sigma_j^2}{\sigma_j^2 - \hat{\sigma}_{i^*}^2}}. \tag{1.3.18}$$

So, this is our result. To prove it, we used the approximation

$$I \approx (I - D_1 E_1)^{-1} \tag{1.3.19}$$

so, we can write

$$\begin{aligned}(I - D_1 E_1)^{-1} =& D_1^{-0.5} \left(I - D_1^{0.5} E_1 D^{0.5}\right)^{-1} D_1^{0.5} \\ =& D_1^{-0.5} (I + D_1^{0.5} E_1 D_1^{0.5} + (D_1^{0.5} E_1 D_1^{0.5})^2 + \ldots) D_1^{0.5}\end{aligned} \tag{1.3.20}$$

where $D_1^{0.5}$ designates the square root of the matrix D_1 obtained by taking all the eigenvalues and replacing them by their square root. We also use the geometric series development for the last equation above:

$$(I - D_1^{0.5} E_1 D^{0.5})^{-1} = I + D_1^{0.5} E_1 D_1^{0.5} + (D_1^{0.5} E_1 D_1^{0.5})^2 + \ldots$$

which is valid as soon as $D_1^{0.5} E_1 D_1^{0.5}$ has all eigenvalues strictly smaller than 1 in absolute value. If all these eigenvalues have their absolute values much smaller than 1, then we can use the approximation:

$$I \approx I + D_1^{0.5} E_1 D_1^{0.5} + (D_1^{0.5} E_1 D_1^{0.5})^2 + \ldots$$

replacing the expression on the right side of the equation above by I this into the right side of 1.3.20, we find:

$$\begin{aligned}& D_1^{-0.5} (I + D_1^{0.5} E_1 D_1^{0.5} + (D_1^{0.5} E_1 D_1^{0.5})^2 + \ldots) D_1^{0.5} \\ & \approx D_1^{-0.5} \cdot I \cdot D_1^{0.5} \\ & = I\end{aligned}$$

which with the help of 1.3.20 leads

$$(I - D_1 E_1)^{-1} \approx I$$

and our 1.3.19. So, this is the last thing remaining to be proven in order to establish 1.3.18. Again, we need $D_1^{0.5} E_1 D_1^{0.5}$ to have spectral norm close to zero, which is the same as looking at the spectral norm of $|D_1|^{0.5} E_1 |D_1|^{0.5}$, where the matrix $|D_1|$is obtained from D_1 by replacing the eigenvalues by their absolute values.

Again E_1 is the matrix obtained from

$$E = \hat{\mathrm{COV}}[\vec{X}] - \mathrm{COV}[\vec{X}]$$

by deleting the first row and column. Similarly we take the diagonal matrix with j-th entry equal to $\sigma_j^2/(\sigma_j^2 - \hat{\sigma}_{i^*2} i^2)$ and then delete the first row and column to obtain D_1 from the finite dimensional approximation. However, we do not attempt to bound $|D_1 E_1|$ in what follows. Rather, we work with bounding the spectral norm of the matrix $|D_1|^{\frac{1}{2}} \cdot E_1 \cdot |D_1|^{\frac{1}{2}}$.

Let us go back to the three dimensional case, which is good to visualise what is going on:

$$
\begin{aligned}
&|D_1|^{\frac{1}{2}} E_1 |D_1|^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{1}{\sqrt{|\sigma_Y^2 - \sigma_X^2 - \Delta\lambda|}} & 0 \\ 0 & \frac{1}{\sqrt{|\sigma_Z^2 - \sigma_X^2 - \Delta\lambda|}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{Y^2} N_{22} & \sigma_Y \sigma_Z N_{23} \\ \sigma_Y \sigma_Z N_{32} & \sigma_{Z^2} N_{33} \end{bmatrix} \\
&\cdot \begin{bmatrix} \frac{1}{\sqrt{|\sigma_Y^2 - \sigma_X^2 - \Delta\lambda|}} & 0 \\ 0 & \frac{1}{\sqrt{|\sigma_Z^2 - \sigma_X^2 - \Delta\lambda|}} \end{bmatrix} \\
&= \frac{1}{\sqrt{n}} \begin{bmatrix} \frac{\sigma_Y}{\sqrt{|\sigma_Y^2 - \sigma_X^2 - \Delta\lambda|}} & 0 \\ 0 & \frac{\sigma_Z}{\sqrt{|\sigma_Z^2 - \sigma_X^2 - \Delta\lambda|}} \end{bmatrix} \cdot \begin{bmatrix} N_{22} & N_{23} \\ N_{32} & N_{33} \end{bmatrix} \\
&\cdot \begin{bmatrix} \frac{\sigma_Y}{\sqrt{|\sigma_Y^2 - \sigma_X^2 - \Delta\lambda|}} & 0 \\ 0 & \frac{\sigma_Z}{\sqrt{|\sigma_Z^2 - \sigma_X^2 - \Delta\lambda|}} \end{bmatrix}
\end{aligned}
$$

Recall that E is the error matrix when estimating the covariance matrix:

$$E = \hat{\mathrm{COV}}[\vec{X}] - \mathrm{COV}[\vec{X}]$$

and E_1 is obtained from E by deleting the first row and columns from E. So, we have E_1 is equal to

$$E_1 = \frac{1}{\sqrt{n}} \begin{bmatrix} \sigma_Y & 0 \\ 0 & \sigma_Z \end{bmatrix} \cdot \begin{bmatrix} N_{22} & N_{23} \\ N_{32} & N_{33} \end{bmatrix} \cdot \begin{bmatrix} \sigma_Y & 0 \\ 0 & \sigma_Z \end{bmatrix} \tag{1.3.21}$$

We note the similarity between the formula for the covariance matrix estimation error given in 1.3.21 and the formula , for

$$|D_1|^{\frac{1}{2}} E_1 |D_1|^{\frac{1}{2}}. \tag{1.3.22}$$

This shows that the matrix 1.3.22 can be interpreted as a covariance estimation matrix, but

with the eigenvalues of the covariance not being to

$$\frac{\sigma_Y}{\sqrt{|\sigma_Y^2 - \sigma_X^2 - \Delta\lambda|}}$$

and

$$\frac{\sigma_Z}{\sqrt{|\sigma_Z^2 - \sigma_X^2 - \Delta\lambda|}}$$

Now, let us go back to the p dimensional case. Similarly, we get that matrix 1.3.22 is the covariance estimation error matrix, when the ground truth eigenvalues are:

$$\frac{\sigma_1^2}{(\sigma_1^2 - \hat{\sigma}_{i^*}^2)^2}, \frac{\sigma_2^2}{(\sigma_2^2 - \hat{\sigma}_{i^*}^2)^2}, \ldots, \frac{\sigma_{i-1}^2}{(\sigma_{i-1}^2 - \hat{\sigma}_{i^*}^2)^2}, \frac{\sigma_{i+1}^2}{(\sigma_{i+1}^2 - \hat{\sigma}_{i^*}^2)^2}, \ldots, \frac{\sigma_p^2}{(\sigma_p^2 - \hat{\sigma}_{i^*}^2)^2}$$

where we act as if $\hat{\sigma}_{i^*}^2$ would be non-random.

We can figure out the spectral norm of $D_i E_i$ up to a universal constant thanks to the break through result of Koltschinksi an Klounici [2]. They show that for an estimated covariance matrix, the spectral norm of the error matrix $|E|$ is typically bounded by

$$|E| \leq C \cdot \left(\max_j \frac{\sigma_j}{\sqrt{n}} \right) \sqrt{\sum_j \sigma_j^2} \tag{1.3.23}$$

where $C > 0$ is a universal constant, which does not depend on n or on the sequence $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$ of ground truth eigenvalues. They get a hard edge sharply exponential decaying property for the probability to have eigenvalues bigger than the bound1.3.23. So, we can apply the formula of koltschinksi and Lounici to our matrix 1.3.22, since that matrix is also covariance error matrix. For this we need to replace σ_j by $\frac{\sigma_j}{\sqrt{|\sigma_j^2 - \hat{\sigma}_{i^*}^2|}}$ for every $j \neq i$ in our bound 1.3.23. This gives us a tied bound, which typically holds for the spectral norm of $|D_i|^{0.5} E_i |D_i|^{0.5}$: So, with high probability:

$$\texttt{spectral norm of } |D_i|^{0.5} E_i |D_i|^{0.5} \leq C \cdot \left(\frac{1}{\sqrt{n}} \max_{j \neq i} \frac{\sigma_j}{\sqrt{|\sigma_j^2 - \hat{\sigma}_{i^*}^2|}} \right) \cdot \sqrt{\sum_{j \neq i} \frac{\sigma_j^2}{|\sigma_j^2 - \hat{\sigma}_{i^*}^2|}}$$

Noting that for fixed i, expression

$$\frac{\sigma_j^2}{|\sigma_j^2 - \sigma_i^2|} = \frac{1}{|1 - (\frac{\sigma_i^2}{\sigma_j^2})|}$$

becomes smaller as σ_j^2 moves away from σ_i^2 in both directions, we get that the maximum is about:

$$\max_{j \neq i} \frac{\sigma_j^2}{|\sigma_j^2 - \sigma_i^2|} \approx \frac{\sigma_i^2}{\texttt{spectral gap}_\texttt{i}}$$

so that we get

$$\texttt{spectral norm of } |D_i|^{0.5} E_i |D_i|^{0.5} \leq C \cdot \frac{\sigma_i}{\sqrt{n}\sqrt{\texttt{spectral gap}_{i^*}}} \cdot \sqrt{\sum_{j \neq i} \frac{\sigma_j^2}{|\sigma_j^2 - \hat{\sigma}_{i^*}^2|}} \tag{1.3.24}$$

So for approximation 1.3.18 to hold up to a smaller error term, we simply need the approximation 1.3.19. On the other hand, for 1.3.19 to hold, we need the spectral norm of $|D_i|^{0.5} E |D_i|^{0.5}$ to be close to 0. To guarantee this, we can use condition 1.3.24. So, we need the right side of 1.3.24 to be quite a bit below 1. Formally we want a small constant ϵ so that $0 < \epsilon < 1$ and 1.3.24 is less than ϵ, which gives the condition on the sample size n:

$$\sqrt{n} \geq C \cdot \frac{\sigma_i}{\epsilon\sqrt{\texttt{spectral gap}_{i^*}}} \cdot \sqrt{\sum_{j \neq i} \frac{\sigma_j^2}{|\sigma_j^2 - \hat{\sigma}_{i^*}^2|}} \tag{1.3.25}$$

This typically will hold, for the eigenvalues of order bigger or equal to $O(p^{\frac{1}{2}})$ assuming the eigenvalues σ_j^2 to be regularly spaced.

So, we assume standardized data, which means $\sum_j \sigma_j^2 = p$. Hence, if a certain type of eigenvalues has a sum less than $O(p)$, they would be not relevant. Hence if we consider eigenvalues of order $O(p^\beta)$ with $0 < \beta < 1$, then there needs to be $O(p^{1-p})$ of them at least, since otherwise their sum would be too small to play an important role. If they are spaced regularly, then the order of the spectral gap must be $O(p^\beta / p^{1-\beta}) = O(p^{2\beta - 1})$.

Now, with enough regularity of the eigenvalues, the expression on the right side of 1.3.25 is approximately equal to $\frac{\sigma_i^2}{\texttt{spectral gap}_i}$. Then plugging in the formula $O(p^{2\beta-1})$ for the spectral gap and $O(p^{\beta})$ for σ_i^2 into 1.3.25, we get that condition that 1.3.25 is satisfied when $\beta > 0.5$

1.4 The Case of large c for the Sample Size $n = c \cdot p$

The current section is for proof of an approximation formula for the difference between the spectrum of sample covariance and ground truth covariance in the case that the constant c is very large. For this we assume as usual a data matrix X of dimension $n \times p$, where $n = c \cdot p$ with i.i.d. normal rows with expectation 0. Then, we let p go to infinity. Our approximation formula is supposed to hold, for large c. Again, let

$$COV[\vec{X}] = \Sigma_p := E[X^t X]$$

denote the $p \times p$ ground truth covariance matrix, which is also denoted by Σ_p. We denote by

$$C\hat{O}V[\vec{X}] = \hat{\Sigma}_p = \frac{X^t \cdot X}{n}$$

the sample covariance matrix. Again, recall that we denote by $\hat{\sigma}_j^2$ the j'th eigenvalue of the sample covariance and by σ_j^2 the j-th eigenvalue of the ground truth covariance. In previous cases, we ordered the eigenvalues in decreasing order. The goal of this section, is to show that the approximation

$$\hat{\sigma}_i^2 - \sigma_i^2 \approx \frac{\sigma_i^2}{n} \sum_{s \notin J_i^k} \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 - \hat{\sigma}_i^2}, \tag{1.4.1}$$

holds given c is sufficiently large. The interval $J_i^k = [i-k, i+k]$ is defined so that the sum approximate the improper integral that is so that:

$$\frac{1}{p}\sum_{s \notin J_i^k} \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 - \hat{\sigma}_i^2} \approx \int \frac{x}{x - \hat{\sigma}_i^2} dF^W(x). \tag{1.4.2}$$

(Note that for i.i.d variables $Y_1, Y_2, \ldots, Y_p$ with a density f_Y, in order to have the approximation

$$\frac{1}{p}\sum_{s} \frac{Y_s}{Y_s - z} \approx \int \frac{x}{x - z} f_y(x) dx$$

we need to remove a few of the Y's closest to z.....) Here are some detailed explanations. Again, c is constant. The idea is that if we take c really large, but then keep it constant whilst p goes to infinity. $\hat{\sigma}_i^2 - \sigma_i^2$ scales like $1/c$. So, we define a to be:

$$a := c \cdot (\hat{\sigma}_i^2 - \sigma_i^2)$$

We view a as a function of σ_i^2 or equivalently as a function of $\hat{\sigma}_i^2$. For c large enough, a should not change a lot and we view it in terms of c as a constant, which depends on eigenvalue we choose. Therefore, the goal of this section is to show that for c large enough, a equals the left side of the approximation 1.4.2 up to a small order term, which would then imply 1.4.1. Instead we are going to prove that a is the right side of 1.4.1 plus a term $O(\frac{1}{C})$ at the limit after p goes to ∞ holding c fixed1.4.44.

Now we assume that we have the data matrix X, which is n times p. For the matrix X, it has the property that all the columns and rows are independent normal random variables with expectation 0. More specifically, we assume that there is a normal random vector of length p with independent entries

$$\vec{X} = (X_1, X_2, \ldots, X_p)$$

where $E(X_j) = 0$ for $j = 1, 2, \ldots, p$ and $X_1, X_2, \ldots, X_p$ are independent. We assume that they are independent because if we would have data with dependent columns, we could just change coordinate system and work with principal components and so get independent coordinates. We assume that $\mathrm{Var}(X_j) = \sigma_j^2$. Hence, the covariance matrix $\mathrm{COV}(\vec{X})$ is a diagonal matrix

$$\mathrm{COV}(\vec{X}) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_3^2 & \ldots & 0 \\ \ldots & & & & \\ 0 & 0 & 0 & \ldots & \sigma_p^2 \end{bmatrix} \tag{1.4.3}$$

Again, we have the $n \times p$ data matrix X:

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \ldots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \ldots & X_{2p} \\ X_{31} & X_{32} & X_{33} & \ldots & X_{3p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \ldots & X_{np} \end{bmatrix} \tag{1.4.4}$$

So, the rows of X are each distributed like $\vec{X}$ and independent of each other. Since $E[\vec{X}] = \vec{0}$, the estimated covariance matrix (sample covariance) is given by

$$\hat{\mathrm{COV}}(\vec{X}) := \frac{X^T X}{n}.$$

Now, we are going to look at the data X without the i-th column. More exactly, we are going to replace the i-th column by zeros, and then compute the sample covariance matrix. This estimated covariance matrix will be denoted by $\hat{\mathrm{COV}}(\vec{X})_{SUB}$. So, we have X_{SUB} is

defined by

$$X_{SUB} := \begin{bmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1(i-1)} & 0 & X_{1(i+1)} & \dots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2(i-1)} & 0 & X_{2(i+1)} & \dots & X_{2p} \\ X_{31} & X_{32} & X_{33} & \dots & X_{3(i-1)} & 0 & X_{3(i+1)} & \dots & X_{3p} \\ \vdots & \vdots & \vdots & \dots & \vdots & 0 & \vdots & \dots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \dots & X_{n(i-1)} & 0 & X_{n(i+1)} & \dots & X_{np} \end{bmatrix}$$

Hence the sample covariance matrix for this "reduced" data matrix is given by:

$$\hat{\mathrm{COV}}(\vec{X})_{SUB} = \frac{X_{SUB}^T X_{SUB}}{n}. \tag{1.4.5}$$

The above estimated covariance matrix has the i-th row and i-th column being 0. Other entries are clearly the same as for the full sample covariance $X^T X/n$. Now, one eigenvalue of the reduced sample covariance 1.4.5 is equal to 0. Others are denoted by

$$\hat{\sigma}^2_{SUB,1} > \hat{\sigma}^2_{SUB,2} > \hat{\sigma}^2_{SUB,3} > \ldots > \hat{\sigma}^2_{SUB,(p-1)}. \tag{1.4.6}$$

The eigenvalues of the original sample covariance $\hat{\mathrm{COV}}[\vec{X}]$ are denoted by:

$$\hat{\sigma}^2_1 > \hat{\sigma}^2_2 > \ldots > \hat{\sigma}^2_p. \tag{1.4.7}$$

In the lemma 1.4.2, we show the interlacing property. That is we always have:

$$\hat{\sigma}^2_1 \geq \hat{\sigma}^2_{SUB,1} \geq \hat{\sigma}^2_2 \geq \hat{\sigma}^2_{SUB,2} \geq \cdots \geq \hat{\sigma}^2_{p-1} \geq \hat{\sigma}^2_{SUB,(p-1)} \geq \hat{\sigma}^2_p$$

Now, we are going to condition on the data without column i, namely condition on X_{SUB}.

Then, the eigenvalues

$$\hat{\sigma}^2_{SUB,1} > \hat{\sigma}^2_{SUB,2} > \hat{\sigma}^2_{SUB,3} > \ldots > \hat{\sigma}^2_{SUB,(p-1)}.$$

are no longer random. When we add the random i-th column to the matrix X_{SUB}, new eigenvalues of the full sample covariance matrix, that is $\hat{\sigma}^2_1 > \hat{\sigma}^2_2 > \ldots > \hat{\sigma}^2_p$, become random. We are going to study the evolution of "this particle process". That is how we get the eigenvalues 1.4.7 from 1.4.6. Now, we denote by ν_j the eigenvalue $\hat{\sigma}^2_{SUB,j}$, for all $j = 1, 2, \ldots, p-1$. Again, we assume that the spectrum of the ground truth covariance $\Sigma_p = COV[\vec{X}]$ converges to a limit with distribution function denoted by F^{Σ} as p goes to infinity. Also, the empirical distribution of the sample covariance matrix 1.4.3 is denoted by F^{W_p}, whilst the spectrum of the restricted sample covaraince $C\hat{O}V[\vec{X}]_{SUB}$ is denoted by $F^{W_{p-1}}$, where we leave out 0. We assume F^{Σ_p} converges and so F^{W_p} must also converge to a limit F^W, so called Whishard distribution. One can, for example, determine eigenvalues for Σ_p by choosing at random i.i.d. from the distribution F^{Σ}, which means that we could have that $\sigma^2_1 > \sigma^2_2 > \ldots > \sigma^2_p$ as a set obtained by choosing p i.i.d. values from the distribution F^{Σ}. Or one could choose in a more regular way to get faster convergence of F^{Σ_p}. Now, in our notation σ^2_i is the one we leave out. When we add σ^2_i to the ground truth spectrum, we go from the empirical distribution $F^{W_{p-1}}$ to F^{W_p}. Since we have convergence of F^{W_p} to F^W, we need

$$G^p := p \cdot F^{W_p} - (p-1) \cdot F^{W_{(p-1)}} \tag{1.4.8}$$

to converge weakly to F^W, at least when the added eigenvalue σ^2_i is chosen at random from the distribution F^{Σ}.

At this stage we are ready to summarize the rest about how we show that our approximation 1.4.1 holds, for large c. Let's look at a few examples first.

EXAMPLE 1: Assume for example $p = 7$, and that we have the spectrum of the

restricted covariance $\hat{\mathrm{COV}}(\vec{X})_{SUB}$ given by:

$$\{\nu_1 = 7, \nu_2 = 6, \nu_3 = 5, \nu_4 = 4, \nu_5 = 3, \nu_6 = 2\}$$

whilst the full sample covariance 's $\hat{Cov}(\vec{X})$ spectrum would be:

$$\{\hat{\sigma}_1^2 = 7, \hat{\sigma}_2^2 = 6, \hat{\sigma}_3^2 = 5, \hat{\sigma}_4^2 = 4.5, \hat{\sigma}_5^2 = 4, \hat{\sigma}_6^2 = 3, \hat{\sigma}_7^2 = 2\}.$$

we see that the difference consists in one point, which has been added. We will denote that point by ξ, so in the current example, we find $\xi = 4.5$. In reality it is unlikely that only one points gets added. So, let us look at a more realistic example.

EXAMPLE 2: Again $p = 7$ and let spectrum of $\hat{\mathrm{COV}}(\vec{X})_{SUB}$ be as before ,but the spectrum of the ground truth be changed to:

$$\{\hat{\sigma}_1^2 = 7, \hat{\sigma}_2^2 = 6, \hat{\sigma}_3^2 = 5.5, \hat{\sigma}_4^2 = 4.5, \hat{\sigma}_5^2 = 3.5, \hat{\sigma}_6^2 = 3, \hat{\sigma}_7^2 = 2\}.$$

In this case, eigenvalues $\nu_1, \nu_2, \nu_5, \nu_6$ are not changed, but all the others are. So we could not view the change as adding one single point. However, we will still do so by viewing the point added ξ to be a random variable with a density function, which is zero outside the interval $[3.5, 5.5]$ and centered maybe, in the current case, at 4.5. Indeed in that interval the total number of points get increased by one when you go from restricted sample covariance matrix spectrum to full sample covariance. There are two approaches presented in our research. One is heuristic and maybe easier to understand. It first shows when we take $C > 0$ really large, we get a situation like the one presented in the current example: most eigenvalues barely change when we add the additional dimension to go from $\hat{\mathrm{COV}}(\vec{X})_{SUB}$ to $\hat{\mathrm{COV}}(\vec{X})$. And the most serious change happens in a restricted interval, which is centered in a certain location. That location could be viewed as the place where we added a point. The heuristic argument is then to say that if the additional eigenvalue σ_i^2 is the i-th eigen-

value of the ground truth spectrum, then this should also add a "point" in the i-th position of the sample covariance. If we assume this to be true, one explains in Section 1.4.2, that this translates into our formula 1.4.1 holding up to a small error term.

Now, the approach we pursue in the rest of this Section is obtained by writing down the equation for the distribution of ξ. Let us see one more example:

EXAMPLE 3: Take the same restricted spectrum as before, but let the spectrum of the full sample covariance be:

$$\{\hat{\sigma}_1^2 = 7.1, \hat{\sigma}_2^2 = 6.1, \hat{\sigma}_3^2 = 5.5, \hat{\sigma}_4^2 = 4.5, \hat{\sigma}_5^2 = 3.5, \hat{\sigma}_6^2 = 2.9, \hat{\sigma}_7^2 = 1.9\}$$

So, this time all the eigenvalues are changed a little bit. However, those further from center are changed much less. So, how could we model this as one point ξ added to the spectrum? The answer is that we take the ratio of how much they get moved to the spectral gap as the probability distribution function. For example, we see, in current example, that between ν_2 and $\hat{\sigma}_2^2$, there is only a distance of 0.1. So we will assume that the random variable ξ, which represents the change in spectrum as one point random variable added, would have a probability of 0.1 to be to the left of ν_2. In other words, we model the probability of ξ by the ratio:

$$P(\xi \leq \nu_j) = E\left[\frac{\hat{\sigma}_j^2 - \nu_j}{\nu_{j-1} - \nu_j}\right] \tag{1.4.9}$$

or we should probably take the expectation on the right side of the equation above. If we take the distribution function G^p as defined in 1.4.8, then at the limit we should get F^W. So a microscopic moving average of G^p should converge to F^W as well. Recall that we had defined a to be

$$a = c \cdot (\hat{\sigma}_i^2 - \sigma_i^2)$$

The goal is to determine a at the limit when p goes to infinity. The way to calculate a is as follows. At the limit we know that ξ must have limit distribution F^W. Recall that we denote by σ_i^2 the eigenvalue of the ground truth covariance matrix. Now, we can add a

value chosen at random among $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{p-1}^2$. In this way, we get a random variable T with probability distribution $F^{\Sigma_{p-1}}$. Then, $\hat{\sigma}_i^2$ is a random variable with distribution F^{W_p}, which we denote by S and we get

$$\sigma_i^2 = T = S - \frac{a(S)}{c}$$

In order to calculate the value of a, what we do in the remainder of this section is simple: since ξ and S are supposed to have the same probability distribution F^W at the limit, we write the equation:

$$P(S \leq x_0) = \int P(\xi \leq x_0 | S = s) dF^W(s). \tag{1.4.10}$$

which is held for every x_0. Now, this is one equation and we have one unknown a. So we can solve for a given a formula for the conditional probability in the integral on the right side of 1.4.10. This formula is obtained from an exact formula 1.4.52 and 1.4.53 for $\hat{\sigma}_j^2 - \sigma_i^2$. This leads to the approximation 1.4.54, which holds up to a small order term. And we can rewrite the approximation as:

$$\frac{c \cdot (\hat{\sigma}_j^2 - \sigma_i^2 + \frac{\sigma_i^2}{c} \sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2})}{\nu_j \cdot \sigma_i^2} \cdot \frac{p}{\nu_{j-1} - \nu_j} \approx (\nu_{j-1} - \nu_j) \cdot \sum_{s \notin J_j^K} \frac{\nu_s \mathcal{N}_s^2}{\nu_s - \hat{\sigma}_j^2} \tag{1.4.11}$$

where $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_{p-1}$ are conditioned on $\nu_1, \ldots, \nu_{p-1}$ i.i.d. standard normal. Also, the interval $J_j^K = [j - K, j + K]$ is to leave out enough uncontrolled small term in the sum on the right of 1.4.11, so as to get the sum to be close to the corresponding indefinite integral.

Similar equation to 1.4.54 is given in [1]. However, the novelty of our research is that we understood this equation and 1.4.11 is not to determine the macroscopical difference between sample spectrum and ground truth. Rather it is to determine the evolution of sample viewed as particle process as we add one additional dimension to the data each time step and observe the resulting evolution. Indeed, by the interlacing property we know

that

$$\hat{\sigma}_j^2 \in [\nu_{j-1}, \nu_j] \tag{1.4.12}$$

so conditioning on $\nu_1, \nu_2, \ldots, \nu_{p-1}$ the macroscopical position of $\hat{\sigma}_j^2$ is no longer to be determined. It is its microscopical relative position within the interval on the right of 1.4.12, which equation 1.4.11 allows to determine. By relative microscopical position we mean: the ratio

$$\frac{\hat{\sigma}_j^2 - \nu_j}{\nu_{j-1} - \nu_j} \tag{1.4.13}$$

Now note that we can solve equation 1.4.11 to determine the value of $\hat{\sigma}_j^2$ inside the interval 1.4.12. Also, note that the left side of 1.4.11 is not affected by the exact position of $\hat{\sigma}_j^2$ inside that interval except for a small order term. Hence, the value of the ratio 1.4.13 can be viewed as the value of a function $g(.)$ of the left side of 1.4.11. The same thing holds when we take the expectation:

$$g\left(\frac{c \cdot (\hat{\sigma}_j^2 - \sigma_i^2 + \frac{\sigma_i^2}{c}\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2})}{\nu_j \cdot \sigma_i^2 \cdot f^W(\nu_j)}\right) = E\left[\frac{\hat{\sigma}_j^2 - \nu_j}{\nu_{j-1} - \nu_j} | S\right], \tag{1.4.14}$$

where we replaced $p/(\nu_{j-1} - \nu_j)$ by the probability density f^W at the limit. We assume that microscopically the adjacent spectral gaps have a joint distribution, which asymptotically does not depend on location or scale once re-scaled by $\nu_j - \nu_{j-1}$. We know that the conditional probability for ξ is less than x_0, $P(\xi \leq x_0|S)$, is given by the expected ratio on the right side of 1.4.14 according to 1.4.9. We can thus replace the conditional probability inside the integral on the right side of 1.4.10 by the expression on the left of 1.4.14. We would then put $\sigma_i^2 = S - a(S)/c$ and put both ν_j and σ_j^2 equal to x_0 in the expression on the left of 1.4.14 and solve. This would work if we could determine the function $g(.)$. Alternatively, for large c we do not need to know everything about $g(.)$. Instead, it is enough to know for large z how $g(.)$ behaves. In the present case, we argue that $g(x) \approx 1/|z|$ as long as z is larger in absolute value than a certain constant.

So, this is the method how to determine a: we take equation 1.4.10 after plug in the formula given for the conditional probability by 1.4.14 and solve for a.

Next we are going to discuss the detail of it. it turns out that for calculation it is easier to do in two steps: first calculate the change in probability when going from S to $T = S - a(S)/c$, then the change in probability from T to ξ. Let us first give one more numerical example, where we can study in details:

EXAMPLE 4: We are dealing with a signed measure. Assume that $p = 6$. And, the eigenvalues are given as follows:

j	ν_j	$\hat{\sigma}_j^2$
1	1	0.9
2	2	1.9
3	3	2.5
4	4	3.3
5	5	4.2
6		5.2

So, note that G^p takes the following values:

$G^p(x)$	0	1	0	1	0	1	0	1
$x \in$	$[-\infty, 0.9)$	$[0.9, 1)$	$[1, 1.9)$	$[1.9, 2)$	$[2, 2.5)$	$[2.5, 3)$	$[3, 3.3)$	$[3.3, 4)$

$G^p(x)$	0	1	0	1
$x \in$	$[4.4.2)$	$[4.2, 5]$	$[5, 5.2)$	$[5.2, \infty]$

First note that due to the interlacing property, we have

$$\hat{\sigma}_1^2 < \nu_1 < \hat{\sigma}_2^2 < \nu_2 \leq \ldots \leq \nu_5 \leq \hat{\sigma}_6^2$$

which implies that the function G^p is alternating between values 0 and 1. The probability distribution function of a random point is increasing and can not be alternatively going up

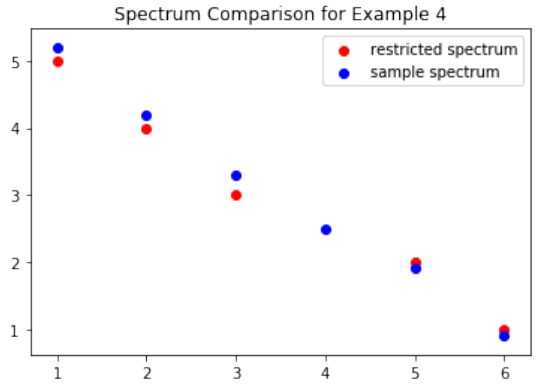


Figure 1.3: Spectrum Comparison for Example 4

and down like G^p does. When we consider our numerical example, we see that ν_1 and $\hat{\sigma}_1^2$ are close to each other. Similarly, in our current example,

$$\nu_1 \approx \hat{\sigma}_1^2, \nu_2 \approx \hat{\sigma}_2^2, \nu_4 \approx \hat{\sigma}_5^2, \nu_5 \approx \hat{\sigma}_6^2 \tag{1.4.15}$$

So, in a very rough approximation we could say that going from spectrum

$$\{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\} \tag{1.4.16}$$

to the spectrum

$$\{\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2, \hat{\sigma}_4^2, \hat{\sigma}_4^2, \hat{\sigma}_5^2, \hat{\sigma}_6^2\} \tag{1.4.17}$$

we "add a point in the area $[\nu_2, \nu_4]$". Now, in the interval $[\nu_1, \nu_2]$ the function G^p is 0 on an sub-interval of length 0.9 and 1 on a sub-interval of length 0.1. So, on average it is 0.1 on that interval. Same thing for the interval $[\nu_2, \nu_3]$. In our example, we can write

$$0.1 = \frac{\hat{\sigma}^2 - \nu_2}{\nu_1 - \nu_2}. \tag{1.4.18}$$

so, if we would do a local smoothing, that is a moving average of G^p. The value between ν_1 and ν_3 would probably be close to 0.1. The moving average happens if we re-simulate the situation many times and then take the average. Now, when we take a moving average of

G^p, we would get those values. Again in reality we are interested in a case with very large p. So instead of $z \mapsto G^p(z)$, we take the map $z \mapsto E[G^p(z)]$, we get a local moving average since we consider values $\nu_1, \nu_2, \ldots, \nu_p$ to be random. But by concentration of measure they fluctuate only microscopically. Hence the moving average will only be microscopical instead of macroscopical. When instead of simulating the data X once, we simulate it many times and then build the average of the function $z \mapsto G^p(z)$. For every simulation we get one realisation of $G^p(z)$. That is for every $z \in \mathbb{R}$, we get a long term average value for $G^p(z)$ denoted by $E[G^p(z)]$. In our example, for z in $[\nu_1, \nu_2]$, $E[G^p(z)]$ would probably be close to 0.1. Of course, we need larger p for this work well.

So here z is non-random. The formula for the value of $E[G^p(z)]$ at $z = \nu_j$ should thus be given by the formula:

$$E[G^p(\nu_j)] = \frac{E[\hat{\sigma}_j^2 - \nu_j]}{E[\nu_{j-1} - \nu_j]} \tag{1.4.19}$$

Let f^{W_p} denote the probability density of the spectrum of the sample covariance. We average over the distribution function F^{W_p}, take the derivative and consider

$$f^{W_p}(z) = \frac{dE[F^{W_p}(z)]}{dz}.$$

We can express the expected distance between eigenvalues in function of the density function f^{W_p}:

$$E[\nu_{j-1} - \nu_j] \approx \frac{1}{p \cdot f_p^W(\nu_j)}. \tag{1.4.20}$$

Recall that by interlacing property, we have that $\hat{\sigma}_j^2$ is in $[\nu_{j-1}, \nu_j]$. The exact location of $\hat{\sigma}_j^2$ is determined by an equation. This equation including the unknown y, in a slightly simplified form, can be written as:

$$c \cdot \frac{h_j}{\nu_j \cdot \sigma_i^2} = -\frac{1}{p}\left(\frac{1}{\nu_{j-1} - y} + \frac{1}{\nu_j - y}\right). \tag{1.4.21}$$

under the constrain $y \in [\nu_{j-1}, \nu_j]$. The value h_j is given and we will look at it later. Now

assume that $h_j > 0$. Then the solution y must be on the right half of the interval $[\nu_{j-1}, \nu_j]$. Assume the length of the interval to be $1/p$, then in that case, the term $-\frac{1}{p}\left(\frac{1}{\nu_{j-1}-y}\right)$ is at most 0.5 in absolute value. So, this leads to the solution of 1.4.21 to satisfy

$$y - \nu_j \approx \frac{\nu_j \cdot \sigma_i^2}{p \cdot h_j \cdot c}, \tag{1.4.22}$$

We can now combine 1.4.22, 1.4.20 and 1.4.19 to obtain:

$$E[G^p(\nu_j)] \approx \frac{\nu_j \cdot f^{W_p} \cdot \sigma_i^2}{c \cdot h_j}. \tag{1.4.23}$$

Again, recall that we take the sample size to be equal to $n = c \cdot p$, where p is the dimension of the space. We take c very, very large but it is a fixed constant, whilst p goes to infinity. The formula for h_j is given as:

$$h_j := \hat{\sigma}_j^2 - \sigma_i^2 + \frac{1}{c} \cdot \frac{\sigma_i^2}{p} \sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2}, \tag{1.4.24}$$

where J_j^K is the integer interval $[j - K, j + K - 1]$ and K is a constant of order $O(1)$. The formula 1.4.21 with our choice of h_j given in 1.4.24 is obtained from 1.4.54, which we prove in the next subsection. We will mention more on that later. Now our goal in this subsection is to show the approximation:

$$\hat{\sigma}_i^2 - \sigma_i^2 \approx -\frac{\sigma_i^2}{n}\left(\sum_{s \notin J_i^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_i^2}\right) \tag{1.4.25}$$

to hold "as well as we want" given c large enough. Note that $\hat{\sigma}_i^2 - \sigma_i^2$ is going to be of order $O(\frac{1}{C})$. So we multiply the left side of 1.4.25 by $1/c$. We define a, the re-scaled difference:

$$a := c \cdot \left(\hat{\sigma}_i^2 - \sigma_i^2\right)$$

So, to say that the approximation 1.4.25 holds as good as we want given c large enough, would mean that the difference between left and right side of 1.4.25 is behaving like $o(\frac{1}{C})$, where a is a $O(1)$ constant different from zero. So, we can treat a like a constant, which is only minimally affected by c, but it depends on σ_i^2.

Now, we can also view $j \mapsto h_j$ as a function of ν_j instead of a function of the index j. This is done by putting

$$h(\nu_j) := h_j.$$

So, let us recapitulate: we add one additional dimension column to the data-matrix X_{SUB}. This means that we add one additional eigenvalue σ_i^2 to the ground truth covariance. The change in spectral distribution due to adding this one dimension is given by the distribution function G^p defined in 1.4.8. The total value of G_p , which is $G^p[(-\infty, \infty)] = 1$. G^p represents a signed measure with positive part having norm p and the negative part having norm $p-1$. Now, $z \mapsto G^p(z)$ is not yet the distribution function of a random variable, since it is not increasing (it represents how all the points in spectrum get change). But we would like to view the change in spectral measure as one point added, which means instead of G^p we would like to have the distribution of one random point, i.e. the distribution function of a random variable. Then we take a local moving average of $G^p(z)$, which corresponds to taking $E[G^p(z)]$. In this way, we obtain a probability distribution function. It means that we can view the change in spectrum as if "one random point ξ was added". We have an exact formula for the probability distribution of ξ. So our point is: there exists a symmetric function $g(.)$ around the origin so that if $x_0 < b_i$ we have:

$$P(\xi < x_0) = E[G^p(x_0)] = g\left(\frac{c \cdot h_j}{\sigma_i^2 \cdot x_0 \cdot f^W(x_0)}\right). \tag{1.4.26}$$

and for $x_0 > b_i$ we get

$$P(\xi > x_0) = E[G^p(x_0)] = g\left(\frac{c \cdot h_j}{\sigma_i^2 \cdot x_0 \cdot f^W(x_0)}\right).$$

b_i is the place where function $z \mapsto h(z)$ is zero. Furthermore for a constant K, we have that if $|z| \geq K$, then $g(z) \approx 1/z$. Now we are going to choose the value for σ_i^2 randomly among all values of $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. In this way $\hat{\sigma}_i$ is also random. Thus $\hat{\sigma}_i$ is chosen randomly among all the eigenvalues of the sample covariance. Hence, it is a random value. It will be denoted by S ;

$$S := \hat{\sigma}_i^2$$

and the random variable S has distribution given by F^{W_p}. With this we find

$$\sigma_i^2 = S - \frac{a(S)}{c}$$

which we define as the variable T, so that

$$T = \sigma_i^2 = S - \frac{a(S)}{c}.$$

Next we look at h_j given in 1.4.24 and find with our notation:

$$h_j = x_0 - T \cdot \left(1 + \frac{1}{c}\Phi(x_0)\right) \tag{1.4.27}$$

where ν_j is denoted by x_0 and the function $\Phi(.)$ is the function

$$\Phi(z) := -\frac{1}{p}\left(\sum_{s \notin [z-\epsilon_z, z+\epsilon_z]} \frac{\nu_s}{\nu_s - z}\right) \approx -\int \frac{s}{s-z} f^{W_p}(s) ds$$

Since now $\hat{\sigma}_i^2$ is random, namely the random variable S, when x_0 is to the left of the 0 of the function $h(x_0, .)$ we can rewrite equation 1.4.26 as

$$P(\xi < x_0 | S) = g\left(\frac{c \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right). \tag{1.4.28}$$

where we replace h_j by the right side of 1.4.27 and where x_0^* is defined by

$$x_0^* := x_0 \left(1 + \frac{\Phi(x_0}{c})\right)^{-1}$$

similarly for x_0 to the right of the zero of h_j, we get

$$P(\xi > x_0|T) = g\left(\frac{c \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right). \tag{1.4.29}$$

Now at the limit S will have as distribution F^W, but the random variable ξ must also have the same distribution at the limit. We can view the process going from S to ξ as a two step process: first we go from S to $T = S - \frac{a(S)}{c}$ Then, we go from T to ξ. Since ξ and S have same distribution, after going to the limit we have that for a fixed non-random x_0, we must have equality

$$P(S \leq x_0) = F_S(x_0) = F_\xi(x_0) = P(\xi \leq x_0)$$

when p goes to infinity. So as we go from S to T then go from T to ξ, the change in the probability distribution function must cancel out, which is

$$F_T(x_0) - F_S(x_0) = -(F_\xi(x_0) - F_T(x_0)) \tag{1.4.30}$$

Now we assume that $s \mapsto a(s)$ is a continuous function. Locally it can be considered like a constant. We also assume it bounded. When we go over from S to $S - \frac{a(S)}{C}$, then locally at x_0, this corresponds to a translation of the probability measure of the random variable S by a small distance $\frac{a(x_0)}{C}$. Now, in a small interval of size Δx, there is a probability mass equalling approximately the size of the small interval Δx times the probability density in that area. So, the amount of probability mass crossing from right to left the point x_0 is approximately $f_S(x_0) \cdot \frac{a(x_0)}{C}$. But since the random variable S at the limit has distribution F^w, we get that

$$F_T(x_0) - F_S(x_0) \approx f^W(x_0)\frac{a(x_0)}{C}. \tag{1.4.31}$$

So what is the change due to going over from the variable T to ξ? Formula 1.4.31 shows that there is long distance mass transportation on a scale $\frac{1}{C}$. In other words, the change $F_\xi(x_0) - F_T(x_0)$ is due to the probability mass, which is to the right of x_0 under the random variable T and gets to the left of x_0 under ξ. Then, there is also mass leaving the interval $[-\infty, x_0]$. That is the probability mass which under T is below x_0 and after is to the right. In other words, we get the formula

$$F_\xi(x_0) - F_T(x_0) = \int_{x_0}^{\infty} P(\xi < x_0 | T = t) f_T(t) dt - \int_{-\infty}^{x_0} P(\xi > x_0 | T = t) f_T(t) dt.$$

We want to replace the conditional probability on the right side of the last equation above using 1.4.28 and 1.4.28. Then we would get an equation, which is not yet quite right:

$$F_\xi(x_0) - F_T(x_0) = \int_{x_0}^{\infty} g\left(\frac{c \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t) dt - \int_{-\infty}^{x_0} g\left(\frac{c \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t) dt \tag{1.4.32}$$

What is the problem with the above? The problem is that our formulas 1.4.28 for $P(\xi \leq x_0 | S)$, we need to have 0 of the function $h(x, S)$ to be to the left of x_0. Otherwise, we get 1 minus the formula. So, we would get

$$P(\xi \leq x_0 | S) = 1 - g\left(\frac{c \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right)$$

For S and T large enough, this is never going to be the case. Recall that $h(x, S)$ is defined as

$$h(x, S) = x - (S - \frac{a(S)}{C}) \cdot \left(1 + \frac{1}{C}\Phi(x)\right)$$

If we set $h(x, S) = 0$, it yields the equation

$$x \cdot (1 + \Phi(x))^{-1} = S - \frac{a(S)}{C} \tag{1.4.33}$$

this would yield a zero as a function of S: $x(S)$. Now, we want to know when that zero of

the function $H(x,S)$ is taken at x_0. We simply replace x by x_0 in the formula 1.4.33 and find

$$s_0 := x^* + \frac{a(S)}{C} \tag{1.4.34}$$

So starting at s_0 the problem starts and goes until $s = x_0$. Except that between x_0 and $x_0 + \frac{a(x_0)}{c}$ the problem is not really there. Because this is the interval, where S is to the right of x_0 but the corresponding T is to the left. So, for our calculation T must go to the right and not jump to the left.

In other words, in order to correct 1.4.32, we need to replace the function $g(.)$ by $1-g(.)$ when S is in the interval $[x_0 + \frac{a(x_)}{C}, x^* + \frac{a(x_0)}{C}]$. This corresponds to the interval from x_0 to x_0^* for T. In other words, since formula 1.4.32 is written with the integrator T we have to replace the function $g(.)$ by $1 - g(.)$ on the interval $[x_0, x_0^*]$. This is the same as change of the integration bound from x_0 to x_0^* in that formula. It yields the correct formula given as:

$$F_\xi(x_0) - F_T(x_0) = \int_{x_0^*}^{\infty} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt - \int_{-\infty}^{x_0^*} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt \tag{1.4.35}$$

Now, we can take advantage of the symmetries of $g(.)$, and evaluate 1.4.35:

$$\int_{x_0^*}^{\infty} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt - \int_{-\infty}^{x_0^*} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt \tag{1.4.36}$$

$$= \int_{x_0^*+\frac{K^*}{C}}^{\infty} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt - \int_{-\infty}^{x_0^*-\frac{K^*}{C}} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt \tag{1.4.37}$$

$$+ \int_{x_0^*}^{x_0+\frac{K^*}{C}} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt - \int_{x_0^*-\frac{K^*}{C}}^{x_0^*} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt \tag{1.4.38}$$

$$\tag{1.4.39}$$

where K^* is a constant, which we take sufficiently large so that expression inside the function $g(.)$ is larger in absolute value than K as long as

$$T \notin [x_0^* - \frac{K^*}{C}, x_0^* + \frac{K^*}{C}] \tag{1.4.40}$$

We can do this because T and $f^W(x_0)$ are supposed to be bounded constants. so, in other words, we have, when 1.4.40 holds,

$$|\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}| \geq K$$

However, recall that K is the constant so that for z with $|z| > K$, we have approximately $g(z) = |1/z|$. Hence, when 1.4.40 holds, we have that

$$g\left(\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) \approx \frac{T \cdot x_0^* \cdot f^W(x_0)}{C \cdot (x_0^* - T)}$$

and hence

$$\begin{aligned}
&\int_{x_0^* + \frac{K^*}{C}}^{\infty} g\left(\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t)dt - \int_{-\infty}^{x_0^* - \frac{K^*}{C}} g\left(\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t)dt \\
&\approx \int_{x_0^* + \frac{K^*}{C}}^{\infty} \left|\frac{T \cdot x_0^* \cdot f^W(x_0)}{C \cdot (x_0^* - T)}\right| f_T(t)dt - \int_{-\infty}^{x_0^* - \frac{K^*}{C}} \left|\frac{T \cdot x_0^* \cdot f^W(x_0)}{C \cdot (x_0^* - T)}\right| f_T(t)dt \\
&= \int_{x_0^* + \frac{K^*}{C}}^{\infty} \frac{T \cdot x_0^* \cdot f^W(x_0)}{C \cdot (T - x_0^*)} f_T(t)dt + \int_{-\infty}^{x_0^* - \frac{K^*}{C}} \frac{T \cdot x_0^* \cdot f^W(x_0)}{C \cdot (T - x_0^*)} f_T(t)dt \\
&\approx \frac{x_0^* f^W(x_0)}{C} \int_{-\infty}^{\infty} \frac{t}{t - x_0^*} dF_T(t)
\end{aligned}$$

which together with 1.4.36 implies

$$\int_{x_0^*}^{\infty} g\left(\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t)dt - \int_{-\infty}^{x_0^*} g\left(\frac{C \cdot (x_0^* - T)}{T \cdot x_0^* \cdot f^W(x_0)}\right) f_T(t)dt \approx \tag{1.4.41}$$

$$\frac{x_0^* f^W(x_0)}{C} \int_{-\infty}^{\infty} \frac{t}{t - x_0^*} dF_T(t) \quad + \quad O(\frac{1}{C^2}) \tag{1.4.42}$$

since

$$\int_{x_0^*}^{x_0+\frac{K^*}{C}} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt - \int_{x_0^*-\frac{K^*}{C}}^{x_0^*} g\left(\frac{C\cdot(x_0^*-T)}{T\cdot x_0^*\cdot f^W(x_0)}\right) f_T(t)dt = O(\frac{1}{C^2}) \tag{1.4.43}$$

due to the symmetry of $g(.)$. To see why the last order above holds, simply replace t in the numerator above by x_0^*. Then by symmetry expression 1.4.43 is exactly zero. Combining 1.4.30, 1.4.31, 1.4.35 and 1.4.41, we find

$$a = -x_0^* \int_{-\infty}^{\infty} \frac{t}{t-x_0^*} dF_T(t) \quad + \quad o(\frac{1}{C})$$

But at the limit as p goes to ∞, we have that T has the distribution of the sample spectrum at the limit. Hence, we can replace F_T by F^W. Furthermore x_0 and x_0^* are at a distance $o(1/C)$ from each other. So replacing x_0^* by x_0 only creates a change of order $o(1/C)$. Hence we get

$$a = -x_0 \int_{-\infty}^{\infty} \frac{t}{t-x_0} dF_T(t) \quad + \quad o(\frac{1}{C}) \tag{1.4.44}$$

which is the main result we want to prove. Or rather, we want the approximation 1.4.1 and instead we proved the version at the limit after p goes to ∞. That version at the limit should imply that the discrete version holds, for p large enough.

1.4.1 Derivation for Main formula about the Effect on Eigenvalues of Adding One Dimension

Now, we are going to change coordinate system. We take the i-th canonical vector $\vec{e}_i$ in $\mathbb{R}^p$. And in the orthogonal complement space to $\vec{e}_i$, we take the principal components of the restricted sample covariance matrix

$$\hat{\mathrm{COV}}[\vec{X}]_{SUB}.$$

In this way, we notice that the matrix $\hat{\mathrm{COV}}[\vec{X}]_{SUB}$ and $\hat{\mathrm{COV}}[\vec{X}]$ are identical except in the i-th column and row. Again, $\hat{\mathrm{COV}}[\vec{X}]_{SUB}$. has its i-th row and column containing only

0's. Since we use the principal components of $\hat{\mathrm{COV}}[\vec{X}]_{SUB}$ as basis, it becomes a diagonal matrix. Let us denote that matrix expressed in that basis by A:

So,

$$A = \begin{bmatrix} \nu_1 & 0 & \dots & 0 & 0 & 0 & & 0 & 0 \\ 0 & \nu_2 & \dots & 0 & 0 & 0 & & 0 & 0 \\ & & & \dots & & & & & \\ 0 & 0 & \dots & \nu_{i-1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \nu_{i+1} & \dots & 0 & 0 \\ & & & \dots & & & & & \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \nu_{p-1} & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{1.4.45}$$

where for simplicity of notation we denote $\hat{\sigma}^2_{SUB,j}$ by ν_j for $\forall j = 1, 2, \cdots, p-1$. So we have $\nu_1 > \nu_2 > \ldots > \nu_p \geq 0$ Then we add a $p \times p$ perturbation matrix E, which is zero everywhere except the i-th row and i-th column to obtain the full sample covariance matrix $\hat{\mathrm{COV}}[\vec{X}]$. That is let:

$$E := \begin{bmatrix} 0 & 0 & \dots & 0 & E_{1i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{2i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{3i} & 0 & \dots & 0 & 0 \\ & & & & \dots & & & & \\ E_{i1} & E_{i2} & \dots & E_{i(i-1)} & E_{ii} & E_{i(i+1)} & \dots & E_{i(p-1)} & E_{ip} \\ & & & & \dots & & & & \\ 0 & 0 & \dots & 0 & E_{(p-2)i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{(p-1)i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{pi} & 0 & \dots & 0 & 0 \end{bmatrix}$$

where E consists of entries of the i-th row and column of the matrix $\hat{\mathrm{COV}}[\vec{X}]$ but expressed

in the basis of the principal components of $\hat{\mathrm{COV}}[\vec{X}]_{SUB}$. So, in that basis, the "full" sample covariance $\hat{\mathrm{COV}}[\vec{X}]$ is equal to $A + E$. Hence, E is the matrix $\hat{\mathrm{COV}}[\vec{X}]$. So we have eigenvalues $\nu_1 > \nu_2 > \ldots > \nu_p \geq 0$ expressed in the basis formed by the principal components of $\hat{\mathrm{COV}}[\vec{X}]_{SUB}$.

Clearly we have that $E_{ij} = E_{ji}$ for $\forall j \in 1, 2, 3, \cdots, p$. Furthermore, in Lemma 1.4.1, we prove that except for E_{ii}, E_{ji} are independent of each other and normal distributed with expectation 0 when conditioning on X_{SUB}, namely conditioning on the whole data except column i. Also, for $j \neq i$, we have that the variance of E_{ij} is equal to $\nu_i \cdot \nu_j / n$. Furthermore $E_{II} \approx \sigma_i^2$.

So we have the diagonal matrix A with elements in the diagonal being $\nu_1 > \cdots > \nu_{p-1}$ and 0. These are also the eigenvalues of A.

Then we add the perturbation E, which only affects the i-the column and row. The new eigenvalues are now $\hat{\sigma}_1 > \ldots > \hat{\sigma}_p$. there is one more. We are going to calculate these new eigenvalues as a function of the E_{ij}'s. To find new eigenvalues we let any of new eigenvalues be denoted by $\lambda + \Delta\lambda$. Therefore, this would be an eigenvalue of $E + A$. Say the corresponding eigenvector is $\vec{\mu} + \Delta\vec{\mu}$, where $\vec{\mu}$ is an eigenvector of A.

With these notations, we have:

$$(A + E)(\vec{\mu} + \Delta\vec{\mu}) = (\lambda + \Delta\lambda)(\vec{\mu} + \Delta\vec{\mu}). \tag{1.4.46}$$

Also, since $\vec{\mu}$ is an eigenvector of A, we have:

$$A\vec{\mu} = \lambda\vec{\mu} \tag{1.4.47}$$

Subtracting equation 1.4.46 from 1.4.47, we find:

$$(A - I\lambda)\Delta\vec{\mu} = -E\vec{\mu} + \Delta\lambda\vec{\mu} + -E\Delta\vec{\mu} + \Delta\lambda\Delta\vec{\mu}. \tag{1.4.48}$$

Now we are going to use 1.4.48 in our case. But to simplify notation, we take $i = 1$ and we take a dimension $p = 3$. The formula we find will be valid in general. Also, without loss of generality, we can take $\Delta\vec{\mu}$ perpendicular to $\vec{\mu}$. In our present case $\vec{\mu} = (1, 0, 0)$ is the first eigenvector of the matrix A, which is equal to

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \nu_1 & 0 \\ 0 & 0 & \nu_2 \end{bmatrix}$$

Since $\Delta\vec{\mu}$ is perpendicular to $\vec{\mu}$, we can write $\Delta\vec{\mu} = (0, \Delta\mu_1, \Delta\mu_2)$ Then, we have the perturbation:

$$E = \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & 0 & 0 \\ E_{31} & 0 & 0 \end{bmatrix} \tag{1.4.49}$$

So, now we can write out equation 1.4.48 with our special case of A and the perturbation matrix E given in 1.4.49 to find:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \nu_1 - \lambda - \Delta\lambda & 0 \\ 0 & 0 & \nu_2 - \lambda - \Delta\lambda \end{bmatrix} \begin{bmatrix} 0 \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix}$$
$$= - \begin{bmatrix} E_{11} \\ E_{21} \\ E_{31} \end{bmatrix} + \begin{bmatrix} \Delta\lambda \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & E_{12} & E_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix}$$

the above equation for matrices can be separated into two parts. The first equation gives us an equation for $\Delta\lambda$:

$$\Delta\lambda = E_{ii} + E_{12}\Delta\mu_2 + E_{13}\Delta\mu_3 \tag{1.4.50}$$

Then the remaining equation can be used to calculate $\Delta\vec{\mu}$ so that gives:

$$\begin{bmatrix} \nu_1 - \lambda - \Delta\lambda & 0 \\ 0 & \nu_2 - \lambda - \Delta\lambda \end{bmatrix} \begin{bmatrix} \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix} = - \begin{bmatrix} E_{12} \\ E_{13} \end{bmatrix}$$

we can solve the above equation for $\Delta\vec{\mu}$ and then plug into equation 1.4.50 to find:

$$\Delta\lambda = E_{11} - \frac{E_{12}^2}{\nu_1 - \lambda - \Delta\lambda} - \frac{E_{13}^2}{\nu_2 - \lambda - \Delta\lambda}$$

So far we have given a three dimensional case. But the last formula above is valid in general and becomes:

$$\Delta\lambda = E_{ii} - \sum_{s=1}^{p-1} \frac{E_{si}^2}{\nu_s - (\lambda + \Delta\lambda)}, \tag{1.4.51}$$

Here λ is the eigenvalue of the restricted covariance matrix A, which is equal to 0. So

$$\lambda = 0$$

Furthermore, $\lambda + \Delta\lambda$ is an eigenvalue of the full sample covariance, which is $A + E = \hat{\mathrm{COV}}[\vec{X}]$. So, in that case $\lambda + \Delta\lambda = \Delta\lambda$ and hence $\Delta\lambda$ represents an eigenvalue of $A + E$. When we consider the equation 1.4.51 as an equation of $\Delta\lambda$ assuming other terms are given, we see that for every interval $[\nu_{s-1}, \nu_s], \forall s = 2, \ldots, p-1$, there is one value inside each interval for $\Delta\lambda$ when solving 1.4.51. This is because RHS of 1.4.51 is strictly decreasing going from ∞ to $-\infty$ as a function of $\Delta\lambda$. So in each interval $[\nu_{s-1}, \nu_s]$ for $s = 2, \cdots, p-1$, there is exactly one solution to 1.4.51, and that solution is the eigenvalue $\hat{\sigma}_s^2$ of the "full" sample covariance matrix. This is another way to prove the interlacing

property proven in Lemma 1.4.2, that is we have

$$\hat{\sigma}_1^2 > \nu_1 > \hat{\sigma}_2^2 > \nu_2 > \ldots > \nu_{p-1} > \hat{\sigma}_p^2$$

where we recall that $\nu_j = \hat{\sigma}_{j,SUB}^2$ is the j-th eigenvalue in decreasing order of the restricted sample covariance $\frac{X_{SUB}^T \cdot X_{SUB}}{n}$. Assume eigenvalues $\nu_1, \nu_2, \ldots, \nu_{p-1}$ of the restricted covariance are given. Then the equation 1.4.51 is the equation, which determines the "dynamix" of the eigenvalues. It shows when we add one eigenvalue in the true covariance matrix, how it is going to affect all the eigenvalues of the sample covariance. We could view this as a particle process, where we add one column after the other to X and have the eigenvalues viewed as particles evolve.

Now, the equation 1.4.51 allows to determine all eigenvalues of the full sample covariance. So for example, the j-th eigenvalue:

$$\hat{\sigma}_j^2 = E_{ii} - \sum_{s=1}^{p-1} \frac{E_{si}^2}{\nu_s - \hat{\sigma}_j^2} \tag{1.4.52}$$

Conditioning on X_{SUB}, which is equivalent to condition on all data columns except the i-th, the term E_{si} for $s \neq i$ are independent normal random variables with variance equalling:

$$\mathrm{Var}[E_{si}] = \frac{\sigma^2 \nu_s}{n} = \frac{\sigma_i^2 \sigma_{SUB,s}^2}{n}$$

Also,

$$E_{ii} = \frac{\sum_j X_{ji}^2}{n} \approx \mathrm{Var}[X_i] = \sigma_i^2$$

Using the last approximation above, we can rewrite equation 1.4.52 as

$$\hat{\sigma}_j^2 - \sigma_i^2 \approx -\frac{\sigma_i^2}{n} \sum_{s=1}^{p-1} \frac{\nu_s \mathcal{N}_s^2}{\nu_s - \hat{\sigma}_j^2} \tag{1.4.53}$$

where $\mathcal{N}_s$ are i.i.d. normal random variables conditioned on X_{SUB}. By the interlacing

property, we have that σ_j^2 is between ν_{j-1} and ν_j. Hence, we are considering the interval with natural number close to j, that is

$$J_j^k := [j - K, j + K]$$

How big K needs to be will be discussed later. We want all the terms ν_t, which are "micoscopically close" to ν_j, to have their indexes in the interval J_j^K. So we can distinguish between terms ν_s close to ν_j (and hence to $\hat{\sigma}_j^2$) and others in equation 1.4.53. For other terms, since terms $\nu_s - \hat{\sigma}_j^2$ are not macroscopically small, we can replace $\mathcal{N}_s^2$ by their expectations 1 and obtain:

$$\sum_{s \notin J_j^K} \frac{\nu_s \mathcal{N}_s^2}{\nu_s - \hat{\sigma}_j^2} \approx \sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2}$$

since the expectation dominates the standard deviation.

Hence we can go back to 1.4.53 to obtain:

$$\hat{\sigma}_j^2 - \sigma_i^2 \approx -\frac{\sigma_i^2}{n} \left(\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} + \sum_{s \in J_j^K} \frac{\nu_s \mathcal{N}_s^2}{\nu_s - \hat{\sigma}_j^2} \right) \tag{1.4.54}$$

Note that for big eigenvalues the term:

$$\hat{\sigma}_j^2 - \sigma_i^2 + \frac{\sigma_i^2}{n} \sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2}$$

is strongly positive. Hence, when we try to solve approximation 1.4.54 with small j, we will have that terms σ_1^2, σ_2^2 are going to be very close to the corresponding ν_s. For large j, when j is closer to p, we have that σ_j^2 is going to be close to ν_{j-1}. When the distance is almost indistinguishable close, we get that basically going from the "particles" $\nu_1, \nu_2, \ldots, \nu_{p-1}$ to the "particles"

$$\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_p^2$$

we leave the ones in the border unchanged and can say that somewhere in the middle there has been a particle added.

When $n = c \times p$, if the constant c is really big, then most particles don't move except in a small interval. Here we make an Ansatz, which later we can at least heuristically justify: in our system we can add any value for σ_i, which is the standard deviation of the column that was left out firstly. The values of X_{SUB} are independent of that value and so are the $\nu_1, \ldots, \nu_{p-1}$. So, we can take any value for σ_i and see what the outcome is.

- Our Ansatz is that (at least when $n = c \times p$ where $c > 0$ is large) we have that the particle added due to adding a column with standard deviation σ_i should be added in the same relative position as is the position of σ_j^2 in the original spectrum.

So, in other words, if σ_i^2 is the i-th eigenvalue of the original spectrum, then the additional eigenvalue added should also be about in the i-position in the sample spectrum, that is to say that we have:

$$\hat{\sigma}_j^2 - \sigma_i^2 + \frac{\sigma_i^2}{n} \sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} \tag{1.4.55}$$

is neither positive nor negative at the point where the particle is added. So according to our Ansats, that is for $j = i$ and hence compared to the other terms in 1.4.54, we would have the term 1.4.56 be small order. So that for $j = i$, we would have:

$$\hat{\sigma}_i^2 - \sigma_i^2 + \frac{\sigma_i^2}{n} \sum_{s \notin J_i^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} \approx 0 \tag{1.4.56}$$

which implies

$$\hat{\sigma}_i^2 - \sigma_i^2 \approx -\frac{\sigma_i^2}{n} \left(\sum_{s \notin J_i^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} \right) \tag{1.4.57}$$

which is the approximation formula we wanted to justify, or rather the continuous version at the limit. For this remember that $\nu_s = \hat{\sigma}_{SUB,s}^2$.

1.4.2 Why Big Constant Makes Particles Being Added Locally

let us consider variables x and y in the following equation:

$$x = -\frac{\sigma_i}{n} \sum_{s \in J_j^K} \frac{\nu_s \mathcal{N}_s^2}{\nu_s - y} \tag{1.4.58}$$

In the above equation we assume all values given except x and y and we further assume the constrain

$$y \in [\nu_{j-1}, \nu_j]. \tag{1.4.59}$$

Note that the function on the RHS of 1.4.58, seen as a function of y, is strictly decreasing going from ∞ to $-\infty$ as y goes from ν_{j-1} to ν_j. So we can write y as $y(x)$ and there is no ambiguity assuming that we know 1.4.59 to hold.

Now take x to be equal to:

$$x = \frac{\nu_{j-1} - \nu_j}{2} - \sigma_i^2 + \frac{\sigma_i^2}{n} \left(\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - (\nu_{j-1} + \nu_j)/2} \right). \tag{1.4.60}$$

Now, the sum

$$\frac{\sigma_i^2}{n} \left(\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} \right) \tag{1.4.61}$$

is not too much affected by the exact value of $\hat{\sigma}_j^2$ since $\hat{\sigma}_j^2$ is contained in the interval $[\nu_{j-1}, \nu_j]$ and in the sum there should be no term close to that interval since we take out all the terms, of which index in J_j^K. That is we take out all elements, which are microscopically close to that interval. So, in the sum 1.4.61, we can replace $\hat{\sigma}_j^2$ by any point in the interval given in 1.4.59 and should only get a small order change. So we can replace $\hat{\sigma}_j^2$ by the middle of the interval, which is $(\nu_j - \nu_{j+1})/2$ and still get a similar value. Hence:

$$\frac{\sigma_i^2}{n} \left(\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - \hat{\sigma}_j^2} \right) \approx \frac{\sigma_i^2}{n} \left(\sum_{s \notin J_j^K} \frac{\nu_s}{\nu_s - (\nu_{j-1} + \nu_j)/2} \right) \tag{1.4.62}$$

Applying 1.4.62 to 1.4.54, we obtain

$$\hat{\sigma}_j^2 - \sigma_i^2 \approx -\frac{\sigma_i^2}{n}\left(\sum_{s\notin J_j^K}\frac{\nu_s}{\nu_s-(\nu_{j-1}+\nu_j)/2}+\sum_{s\in J_j^K}\frac{\nu_s\mathcal{N}_s^2}{\nu_s-\hat{\sigma}_j^2}\right) \tag{1.4.63}$$

When we replace $\hat{\sigma}_j^2$ by $(\nu_{j-1}-\nu_j)/2$ on the right side of the above approximation, we have:

$$x \approx -\frac{\sigma_i}{n}\sum_{s\in J_j^K}\frac{\nu_s\mathcal{N}_s^2}{\nu_s-\hat{\sigma}_j^2}$$

The last approximation above shows that we can determine the value of $\hat{\sigma}_j^2$ up to a small error term by solving equation 1.4.58 for y under the constrain 1.4.59 and where x is defined in 1.4.60.

When we consider equation 1.4.58 with the constrain 1.4.59, then: when x is very negative (large absolute value, but negative), then $y(x)$ is close to ν_j. On the opposite when x in 1.4.58 is very large positive and condition 1.4.59 holds, then $y(x)$ is close to ν_{j-1}.

Now, assume that i is somewhere in the middle of the spectrum. Then for $j << i$ we have that x (as given in 1.4.60) is positive and for $j >> i$ we get that x is negative. In order to understand ,let us consider the following: we assume that $n = C \cdot p$ and the constant C is sufficiently large. Then there is not a big difference between sample spectrum and population spectrum. The difference is still of order $O(1))$ but has a small constant in front. So, in the first approximation x is about $\sigma_j^2 - \sigma_i^2$, which obviously is positive for $j < i$ and negative for $j > i$.

Next we want to see when $j << i$, if x is "big enough" to make the solution y of equation 1.4.58 much closer to ν_j. Because in that case, we get that $\hat{\sigma}_j^2$ can be approximately found using equation 1.4.58 and that $\hat{\sigma}_j^2$ is also going to be very close to ν_j. So the point ν_j will be quite indistinguishable of $\hat{\sigma}_j^2$ for $j << i$. We want to prove the opposite, when $j >> i$, that x is negative enough so that the solution of equation 1.4.58 is close to ν_{j-1}. This would then imply that $\hat{\sigma}_j^2$ would be very close to ν_{j-1}. So, in other words, if we can

prove these two things: x gets negative enough for $j << i$ and positive enough fro $j >> i$, then when we go from the spectrum

$$\nu_1 > \nu_2 > \ldots > \nu_{p-1} \tag{1.4.64}$$

to

$$\hat{\sigma}_1^2 > \hat{\sigma}_2^2 > \ldots > \hat{\sigma}_p^2, \tag{1.4.65}$$

that in principle for indexes away from i, the eigenvalues don't change too much. In that case, we can view the effect of going from 1.4.64 to 1.4.65 as "adding a particle somewhere in the vicinity of ν_i".

What we need for this to work is a regularity of the particles given in 1.4.64. More specifically, assume that:

$$\nu_{j-1} - \nu_j \geq K \times \frac{1}{p} \tag{1.4.66}$$

where $K > 0$ is a constant, the interval J_j^k contains only two integers $J^K = [j-1, j]$. With this we can now rewrite 1.4.58 as

$$c \cdot x = -\frac{\sigma_i}{p} \left(\frac{\nu_{j-1} \mathcal{N}_{j-1}^2}{\nu_{j-1} - y} + \frac{\nu_j \mathcal{N}_j^2}{\nu_j - y} \right) \tag{1.4.67}$$

where we also use that $n = C\dot{p}$, where $C > 0$ is a constant.

Now we want y, the solution of 1.4.72, to be close to ν_j. What do we mean by this? We may, for example, request that $|y - \nu_j|$ is at least 10 times smaller than $\nu_{j-1} - \nu_j$, which means:

$$\frac{|y - \nu_j|}{\nu_{j-1} - \nu_j} \leq \frac{1}{10}$$

and hence with the help of 1.4.66 we find

$$\frac{1}{|y - \nu_j| \cdot p} \geq \frac{10}{K}$$

which with the help of 1.4.72 we obtain as long as

$$C \cdot x \geq \frac{10}{K} \cdot \frac{\nu_j \mathcal{N}_j^2}{\sigma_i^2}$$

The inequality above holds with high probability by simply taking the constant $C > 0$ large enough since σ_i^2, ν_j and K are all of order $O(1)$ and as long as x is not infinitesimal but of order $O(1)$.

Now, say we want for a large (but constant number) l, the solution y of equation 1.4.72 to be l times closer to ν_j than ν_{i-1}. Note that this closeness follows from 1.4.66, 1.4.72 and

$$c \cdot x \geq \frac{l}{K} \cdot \frac{\nu_j \mathcal{N}_j^2}{\sigma_i^2} \tag{1.4.68}$$

In other words,

$$\frac{|y - \nu_j|}{\nu_{j-1} - \nu_j} \leq \frac{1}{l} \tag{1.4.69}$$

follows from 1.4.66, 1.4.72 and 1.4.68. Now, $\hat{\sigma}_j^2$ is the value for y solving 1.4.72 with the contains $y \in [\nu_{j-1} - \nu_j]$. So, if we take l really large (but constant, think of a million for example), then y becomes almost indistinguishable from ν_j. This means that for practical purpose, $\hat{\sigma}_j^2$ and $\nu_j = \hat{\sigma}_{SUB,j}^2$ will be indistinguishable. This is for $j << i$. Similarly for $j >> i$, we can get that $\hat{\sigma}_j^2$ and $\nu_{j-1} = \hat{\sigma}_{SUB,j-1}^2$ will be practically indistinguishable. This means that between the sample covariance and the restricted sample covariance, the difference is mainly in the eigenvalues around the i-th, when we add one eigenvalue of size σ_i^2. Now we need this result to hold uniformly over $i \in 1, 2, \ldots, p$. And we also need this to hold when j is sufficiently close to i. What we want is to obtain that if we take the constant C very big, we get that the effect of adding one additional dimension, for practical purposes, does not change the spectrum except in a narrow region of the spectrum around the i-th eigenvalue.

As long as x is of $O(1)$, we can obtain this by simple taking the constant C in 1.4.68 large enough. So we need x to be bounded below as long as i and j are not too close.

1.4.3 Lemma

In this section we will introduce some lemma that will be used in the following proof.

The first lemma shows the distribution of the restricted covariance matrix. Recall that X is an $n \times p$ matrix with independent columns, where entries in column j have standard deviation σ_j. In order to compute the restricted covariance matrix, firstly we replace the i-th column in X by 0. We denote the new matrix with a zero column as X_{SUB} and the estimated covariance matrix is

$$\hat{\mathrm{COV}}(X)_{SUB} = \frac{X_{SUB}^T \cdot X_{SUB}}{n}. \tag{1.4.70}$$

Then we express our restricted covariance matrix 1.4.70 in the basis of its principal components. Note that the i-th canonical vector

$$(0, 0, 0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^p$$

is a principal component of 1.4.70, which has a 1 in its i-th entries and 0's everywhere else. It is the principal component with corresponding eigenvalue 0. This is because the matrix 1.4.70 has its i-column and i-th row equal to 0. The principal components are simply eigenvectors by the definition of the principal components. Therefore, when you express a matrix in the basis of its principal components, the matrix becomes diagonal with eigenvalues along the diagonal. In the present case, eigenvalues are denoted by $\hat{\sigma}^2_{SUB,j}$ and also denoted as $\nu_j = \hat{\sigma}^2_{SUB,j}$. So, we are going to represent the full covariance matrix in the basis using principal components of the sub-matrix $\mathrm{COV}(X)_{SUB}$. The sub-matrix part gets diagonalized in that basis. Except for the i-th column and i-th row, we are dealing with a diagonal matrix. This is to say that in that basis of eigenvectors of 1.4.70, the full

covariance matrix $\frac{X^T X}{n}$ will take the following form:

$$\begin{bmatrix} \nu_1 & 0 & \dots & 0 & E_{1i} & 0 & \dots & 0 & 0 \\ 0 & \nu_2 & \dots & 0 & E_{2i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{3i} & 0 & \dots & 0 & 0 \\ & & & & \dots & & & & \\ E_{i1} & E_{i2} & \dots & E_{i(i-1)} & E_{ii} & E_{i(i+1)} & \dots & E_{i(p-1)} & E_{ip} \\ & & & & \dots & & & & \\ 0 & 0 & \dots & 0 & E_{(p-2)i} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & E_{(p-1)i} & 0 & \dots & \nu_{p-2} & 0 \\ 0 & 0 & \dots & 0 & E_{pi} & 0 & \dots & 0 & \nu_{p-1} \end{bmatrix} \tag{1.4.71}$$

The next lemma shows that the non-diagonal entries, that is E_{si} for $s \neq i$ are independent joint normal distributed with given variance.

Lemma 1.4.1. *Assume that we express the sample covariance matrix $\frac{X^T X}{n}$ in the basis using principal components of the restricted matrix $\frac{X_{SUB}^T X_{SUB}}{n}$ to obtain a matrix given in 1.4.71. Then, conditioning on X_{SUB}, we have that E_{ij} for $j \neq i$ are independent normal distributed with*

$$\mathrm{Var}[E_{ij}] = \frac{\nu_j \cdot \sigma_i^2}{n} \tag{1.4.72}$$

Proof. Now, assume given an i.i.d. sequence of normal random variables

$$N_1, N_2, \dots, N_{p-1}$$

with expectation 0 and standard deviation σ. Let

$$Y := \sum_j a_j N_j$$

and

$$Z = \sum_j b_j N_j$$

where the a_j's and the b_j's are non-random coefficients. Then, Y and Z are jointly normal with covariance:

$$\mathrm{COV}(Y, Z) = \sigma^2 \sum_j a_j b_j. \tag{1.4.73}$$

Now, let us look at the sample covariance matrix $\frac{X^T X}{n}$ before the change of basis. If we look at the entry in the i-th row and s-th column and denote it by E^*_{is} for $s \neq i$. The entry is the product of i-th column and s-th column of matrix X and divided by n. We condition on X_{SUB}, which means we condition i-th column is a column of i.i.d normal random variables with standard deviation σ_i. In this situation, we can conclude that the entry E^*_{is} for $s \neq i$ are jointly normal distributed conditioned on X_{SUB}. Because these entries are the results of dot product between a vector of coefficients and a vector of i.i.d normal random variables. The vector of i.i.d normal random variables is the i-th column of X. According to formula 1.4.73, in order to find the covariance

$$\mathrm{COV}(E_{is}, E_{it})$$

we need to take the dot product between coefficient vectors. Here E_{si} is the dot product of the s-th column of X(coefficient vector) and the i-th column of X(random variables) and E_{it} is the dot product of the t-th column of X(coefficient vector) and the i-th column of X(random variables). So the covariance $\mathrm{COV}(E_{is}, E_{it}$) is the product of the s-th column times the t-th column times σ_i^2 divided by n^2. But this is the s, t-th entry of the sample covariance times σ_i^2/n. In other words conditioning on X_{SUB}, the coefficients E_{is} for $s \neq i$, are jointly normal distributed with their covariance matrix equal to product of sample covariance matrix and coefficient σ_i^2/n . Now, when you change for a normal vector the basis and take the principal component as a basis, you get a normal vector with independent

components and where the variance of the components are the eigenvalues of the original covariance matrix. In our case, these variances are $\nu_s \frac{\sigma_i^2}{n}$ which proves 1.4.72 □

In the end, we would like to introduce another lemma that will be helpful for the proof. It can be derived from Cauchy Interfacing Theorem and illustrates the relationship between eigenvalues of full sample matrix and eigenvalues of its sub-matrix.

Lemma 1.4.2. *Assume we have full sample covariance matrix X and its sub-matrix defined in the previous part X_{SUB}. There always exists an orthogonal projection P such that:*

$$P^* \times X \times P = X_{SUB}$$

If we let $\sigma_{SUB,j}$ represent the j-th eigenvalue of X_{SUB} and let σ_j represent the j-th eigenvalue of full sample covariance matrix X. Also we assume all eigenvalues are sorted in descending order, which means:

$$\sigma_1 > \sigma_2 > \cdots > \sigma_n$$

and

$$\sigma_{SUB,1} > \sigma_{SUB,2} > \cdots > \sigma_{SUB,n}$$

Then we have the interlacing property:

$$\sigma_j > \sigma_{SUB,j} > \sigma_{j+1}, \forall j \in 1, 2, \cdots, n-1$$

Proof. Without loss of generality, we may assume matrix X is a $n \times n$ matrix and we get X_{SUB} by deleting the last column and last row of matrix X.

Now we consider a $n \times n-1$ projection matrix P as following:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{1.4.74}$$

Then we have:

$$P^* \times X \times P = X_{SUB}$$

Next, we apply Cauchy interlacing theorem directly and have:

$$\sigma_j > \sigma_{SUB,j} > \sigma_{j+1}, \forall j \in 1, 2, \cdots, n-1$$

□

CHAPTER 2
IMPROVED TEXT CLASSIFICATION METHODS BASED ON NAIVE BAYES MODEL

2.1 Improved Naive Bayes with Optimal Correlation Factor for Text Classification

2.1.1 Introduction

In recent years, rapid growth of text documents on the Internet and digital libraries has enhanced the importance of text classification, whose goal is to find categories of each document given their contents. Text classification has many applications in natural language processing, such as topic detection [29], spam filtering [30, 31, 32], author identification [33], web page classification [34] and sentiment analysis [35]. Despite intensive research, it still remains an open problem today.

Although text classification can be realized with schemes having different settings, the fundamental scheme usually consists of two stages: feature generation and classification. In the classification step, there are two optional steps that would benefit the model: feature extraction and feature selection. Many research projects have been done on feature extraction and selection areas, such as some novel feature selection methods proposed by [36, 37, 38]. Other research projects [39] propose a simple heuristic solution of applying a hierarchical tree to assign components to classes, which performs better on large data sets.

For the second stage, classification, it has been studied from both supervised classification and unsupervised clustering. For supervised classification, if we assume all the categories follow independent multinomial distribution and each document is a sample generated by the distribution, a straight-forward idea would be applying Naive Bayes (NB) [40, 41, 42, 43], which uses scores based on the 'probabilities' of each document conditioned on the categories. NB classifier learns from training data to estimate the distribution of

each category, then computes the conditional probability of each document given the class label by applying Bayes rule. The prediction of the class is done by choosing the highest posterior probability.Another important method is Support Vector Machine [44, 45], which is used to find the maximum-margin hyper-plane that divides the documents with different labels.Usually in Support Vector Machine we will label document as 1 and -1. Therefore, it is widely used in binary classification problems. However, Support Vector Machine can also be used for multi-classification by using one-vs-all technique. Apart from Support Vector Machine, we have logistic regression and random forest, which also work well for classification problem. The choice of different models depends on the size of data, problem requirement and interpretation requirement. When we take into account more factors, such as order of the sequence and meaning of words given a large enough data set, we can use deep learning models such as convolution neural network, recurrent neural network [46, 47] like Recurrent Gated Unit and Long Short Term Memory.

For unsupervised problems, [48] proposed SVD (Singular Value Decomposition) for dimension reduction, then use general clustering algorithm such as K-means and K Nearest Neighborhood. There also exist some algorithms based on EM algorithm, such as pLSA (Probabilistic latent semantic analysis)[49], which considers the probability of each co-occurrence of words and documents as a mixture of conditionally independent multinomial distributions. The parameters in pLSA can not be derived, therefore they used the standard EM algorithm for estimation. Using the same idea, but assuming that the topic distribution has sparse Dirichlet prior, [50] proposed LDA (Latent Dirichlet allocation). The sparse Dirichlet priors encode the intuition that documents cover only a small set of topics and topics frequently use only a small set of words. In practice, this results in a better disambiguation of words and a more precise assignment of documents to topics.

Our research focuses on the performance of Naive Bayes approach for the text classification problems. There are multiple reasons: first of all, classification results of Naive Bayes can be easily interpreted using probability, which is more friendly for people to un-

derstand; secondly, Naive Bayes approach works well with small data set compared to neural network and support vector machine. Although it has many advantages, it still requires plenty of well-labelled data for training purpose. Moreover, its conditional independence assumption is rarely held in reality.

Many researcher have studied Naive Bayes for text classification problems. In [51], authors propose a simple, efficient, and effective feature weighting approach, called deep feature weighting (DFW), which estimates the conditional probabilities of naive Bayes by deeply computing feature weighted frequencies from training data. They incorporate feature weighting idea to conditional probability estimates instead of classification of formula. Therefore, the classification formula is defined as follows:

$$c(x) = \arg\max_{c \in C} P(c) \prod_{i=1}^{m} P(\alpha_i | c, W_i)^{W_i}$$

where W_i is the weight of i-th feature. And in [52], authors propose two novel approached, which also incorporate feature weighting idea. First of all, they adapt gain ratio-based feature weighting to Naive Bayes classifier and address the issue on how to define the gain ratio of each feature (word) partitioning a collection of training documents since this method has already been studied on standard Naive Bayes. Moreover, they adapt the decision tree-based feature weighting method to Naive Bayes classifier.

In [53], authors propose a locally weighted naive Bayes text classifiers. [54] proposed a method that finds better estimation of centroid, which helps improve the accuracy of Naive Bayes estimation. An instance-weighting approach was proposed in [55]. In [42] authors propose a new, very simple semi-supervised extension of multinomial Naive Bayes, called Semi-supervised Frequency Estimate (SFE). Their approach learns $\hat{P}(w_i|c)$ in the following way:

$$\hat{P}(w_i|c) = \frac{\sum_{t=1}^{|T_u} f_i^t \hat{P}(c|w_i)_l}{\sum_{j=1}^{|V|} \sum_{t=1}^{|T_u} f_j^t \hat{P}(c|w_j)_l}$$

In [56] authors propose a latent selection augmented naive (LSAN) Bayes classifier. By

introducing a latent feature selection indicator, the global selection index can be factorized into local selection index, which can be calculated explicitly. Then the feature subset selection models can be pruned by thresholding the local selection index and will be used in the classification model.

Although many researchers have studied the Naive Bayes classifier, we focus on the situation where there does not exist enough labelled data for each class. Different from other feature weighting approaches and adaptive approaches, the key part of our approach is the correlation factor. Our motivation is that we believe for each single labelled text data, even if it is single labelled, it may still include information from other labels, which makes it being related to other labels at the same time. Our correlation factor would combine more feature information taken from different classes.

2.1.2 General Setting

Consider a classification problem with sample $x \in S$ and class set C, where

$$C = \{C_1, C_2, ..., C_k\}.$$

We are interested in finding our estimator:

$$\hat{y} = f(x; \theta) = (f_1(x; \theta), f_2(x; \theta), ..., f_k(x; \theta))$$

for y

Then without further notification:

- Assume that all the categories are independent multinomial distributions and each document is a sample independently generated by a certain distribution
- S is the document set and assume the class set C has k different categories: $\{C_1, C_2, ..., C_k\}$.
- For each category C_i, the centroid $\theta_i = (\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_v})$ and θ_i satisfies: $\sum_{j=1}^{v} \theta_{i_j} = 1$

- Assume we have totally v different words, thus for each document $d \in S$: $d = \{x_1, x_2, \cdots, x_v\}$,where x_i represents the number of occurrence for i-th word with $\sum_{j=1}^{v} x_j = m$

- Assume label vector $y = (y_1, y_2, \cdots, y_k)$. For document d in class C_i, $y_i(d) = 1$ and $\sum_{i=1}^{k} y_i = 1$.

- $\hat{y}(d) = f(d;\theta) = (f_1(d;\theta), f_2(d;\theta), ..., f_k(d;\theta))$ is our estimator for y, where θ is the parameter matrix and $f_i(d;\theta)$ is the likelihood function of document d in class C_i.

2.1.3 Naive Bayes classifier in text classification problem

In this section we will discuss the properties of Naive Bayes estimator. Naive Bayes classifier are a group of probabilistic classifiers based on Bayes Theorem. While the term "Naive" is related to assumptions for this classifier, people assume that each feature makes an independent and equal contribution to the classification outcome. For example, if we consider a text classification problem, then each topic like finance, politics, sports and entertainment are considered independent of each other. In the following case, we show that this independence assumption is used to rewrite the probability of an intersection as a product of several separate probabilities. And our research actually tries to relax this assumption. For the other assumption, approaches like feature weighting did pretty well.

In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. Let class C_i $(1 \leq i \leq k)$ with centroid $\theta_i = (\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_v})$ and θ_i satisfies: $\sum_{j=1}^{v} \theta_{i_j} = 1$. Assuming independence of the words, the most likely class

for a document d is computed as:

$$\begin{aligned} label(d) &= P(C_k|d) \\ &\propto P(C_k, d) \\ &= P(d, C_k) \\ &= \operatorname*{argmax}_i P(C_i)P(d|C_i) \\ &= \operatorname*{argmax}_i P(C_i) \prod_{j=1}^{v} (\theta_{i_j})^{x_j} \\ &= \operatorname*{argmax}_i \log P(C_i) + \sum_{j=1}^{v} x_j \log \theta_{i_j}. \end{aligned}$$

This gives the classification criteria once θ is estimated, namely finding the largest among

$$\log f_i(d;\theta) = \log P(C_i) + \sum_{j=1}^{v} x_j \log \theta_{i_j} \quad 1 \leq i \leq k$$

Now we shall derive an maximum likelihood estimator for θ. For a class C_i, we have the standard likelihood function:

$$\begin{aligned} L(C_i, \theta) &= \prod_{d \in S} f_i(d;\theta)^{y_i(d)} \\ &= \prod_{d \in C_i} \prod_{j=1}^{v} \theta_{i_j}^{x_j} \end{aligned} \tag{2.1.1}$$

Take logarithm for both sides, we obtain the log-likelihood function:

$$\log L(C_i, \theta) = \sum_{d \in C_i} \sum_{j=1}^{v} x_j \log \theta_{i_j}. \tag{2.1.2}$$

We would like to solve optimization problem:

$$\begin{aligned} \max \quad & \log L(C_i, \theta) \\ \text{subject to :} \quad & \sum_{j=1}^{v} \theta_{i_j} = 1 \\ & \theta_{i_j} \geq 0 \end{aligned} \tag{2.1.3}$$

The problem (2.1.3) can be explicitly solved by Lagrange Multiplier, for class C_i, we have $\theta_i = \{\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_v}\}$, where:

$$\hat{\theta}_{i_j} = \frac{\sum_{d \in C_i} x_j}{\sum_{d \in C_i} \sum_{j=1}^{v} x_j}. \tag{2.1.4}$$

For estimator $\hat{\theta}$, we have following theorem.

Theorem 2.1.1. *Assume we have normalized length of each document, that is:* $\sum_{j=1}^{v} x_j = m$ *for all documents* $d \in S$*, the estimator* (2.1.4) *satisfies following properties:*

1. $\hat{\theta}_{i_j}$ *is unbiased.*
2. $E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = \frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m}$.

Proof. With assumption $\sum_{j=1}^{v} x_j = m$, we can rewrite (2.1.4) as:

$$\hat{\theta}_{i_j} = \frac{\sum_{d \in C_i} x_j}{\sum_{d \in C_i} m} = \frac{\sum_{d \in C_i} x_j}{|C_i|m}.$$

Since $d = (x_1, x_2, ..., x_v)$ is multinomial distribution in class C_i, we have: $E[x_j] = m\theta_{i_j}$, and $E[x_j^2] = m\theta_{i_j}(1 - \theta_{i_j} + m\theta_{i_j})$.

1.

$$\begin{aligned}
&E[\hat{\theta}_{i_j}] \\
=&E[\frac{\sum_{d\in C_i} x_j}{|C_i|m}] \\
=&\frac{\sum_{d\in C_i} E[x_j]}{|C_i|m} \\
=&\frac{\sum_{d\in C_i} m\theta_{i_j}}{|C_i|m} \\
=&\theta_{i_j}.
\end{aligned}$$

Thus $\hat{\theta}_{i_j}$ is unbiased.

2. By (1), we have:

$$E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = E[\hat{\theta}_{i_j}^2] - 2\theta_{i_j}E[\hat{\theta}_{i_j}] + \theta_{i_j}^2 = E[\hat{\theta}_{i_j}^2] - \theta_{i_j}^2.$$

Then notice

$$\hat{\theta}_{i_j}^2 = \frac{(\sum_{d\in C_i} x_j)^2}{|C_i|^2m^2} = \frac{\sum_{d\in C_i} x_j^2 + \sum_{d\neq d'\in C_i} x_j^d x_j^{d'}}{|C_i|^2m^2}, \tag{2.1.5}$$

where $d = (x_1^d, x_2^d, ..., x_v^d)$.

Since:

$$\begin{aligned}
&E[\frac{\sum_{d\in C_i} x_j^2}{|C_i|^2m^2}] \\
=&\frac{|C_i|m\theta_{i_j}(1-\theta_{i_j}+m\theta_{i_j})}{|C_i|^2m^2} \\
=&\frac{\theta_{i_j}(1-\theta_{i_j}+m\theta_{i_j})}{|C_i|m}
\end{aligned}$$

and

$$
\begin{aligned}
& E[\frac{\sum_{d \neq d' \in C_i} x_j^d x_j^{d'}}{|C_i|^2 m^2}] \\
=& \frac{|C_i|(|C_i|-1)m^2\theta_{i_j}^2}{|C_i|^2 m^2} \\
=& \frac{(|C_i|-1)\theta_{i_j}^2}{|C_i|}
\end{aligned}
$$

Plugging them into (2.1.5) obtains:

$$E[\hat{\theta}_{i_j}^2] = \frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m} + \theta_{i_j}^2,$$

thus: $E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = \frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m}$.

□

The Naive Bayes with multinomial prior distribution has a strong assumption about the data: it assumes that words in documents are independent. However, this assumption clearly does not hold in real world text. There are many different kinds of dependence between words induced by semantic, pragmatic, and conversational structure of a text. Although it has its advantages in practice compared to some more sophisticated models, we propose a new method based on Naive Bayes model that has a better performance by introducing a correlation factor, especially for the situation where there is no sufficient data compared with large number of classes.

2.1.4 Naive Bayes with correlation factor

From Theorem.2.1.1, we can see that traditional Naive Bayes estimator $\hat{\theta}$ is an unbiased estimator with variance $O(\frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m})$. Now we will try to find an estimator, and prove that it can perform better than traditional Naive Bayes estimator.

There are two main approaches to improve the Naive Bayes model: modifying the

feature and modifying the model. Many researchers have proposed approaches to modify the document representation in order to better fit the assumption made by Naive Bayes. These include extracting more complex features such as syntactic or statistical phrases [57], extracting features using word clustering [58] and exploiting relations using lexical resources [59]. We propose an approach that modifies the probabilistic model. Therefore, our model should work well with other document representation modifications to achieve a better result.

Our basic idea is that, even for a single labeling problem, a document d usually contains words appearing in different classes, thus it should include some information from different classes. However, our label y in training set does not reflect that information because only one component of y is 1 and all others are 0. We would like to replace y by $y + t$ in Naive Bayes likelihood function 2.1.1 with some optimized t to get our new "likelihood function" L_1:

$$\begin{aligned} L_1(C_i, \theta) &= \prod_{d \in S} f_i(d; \theta)^{y_i(d)+t} \\ &= \prod_{d \in S} (\prod_{j=1}^{v} \theta_{i_j}^{x_j})^{y_i(d)+t}. \end{aligned} \tag{2.1.6}$$

By introducing the correlation factor t, we include more information between the document and classes, which improves the classification accuracy.

Notice that to compute L_1 of a given class C_i in our estimator, instead of just using documents in C_1 as Naive Bayes estimator, we will use every $d \in S$.

Take logarithm for both sides of 2.1.6, we obtain the log-likelihood function:

$$\log L_1(C_i, \theta) = \sum_{d \in S} \left[(y_i(d) + t) \sum_{j=1}^{v} x_j \log \theta_{i_j} \right]. \tag{2.1.7}$$

Similar to Naive Bayes estimator, we would like to solve optimization problem:

$$\begin{aligned} \max \quad & \log L_1(C_i, \theta) \\ \text{subject to :} \quad & \sum_{j=1}^{v} \theta_{i_j} = 1 \\ & \theta_{i_j} \geq 0 \end{aligned} \tag{2.1.8}$$

Let:

$$G_i = 1 - \sum_{j=1}^{v} \theta_{i_j},$$

by Lagrange multiplier, we have:

$$\begin{cases} \dfrac{\partial \log(L_1)}{\partial \theta_{i_j}} + \lambda_i \dfrac{\partial G_i}{\partial \theta_{i_j}} = 0 \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases}$$

plug in, we obtain:

$$\begin{cases} \displaystyle\sum_{d \in S} \frac{(y_i(d) + t)x_j}{\theta_{i_j}} - \lambda_i = 0, \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases} \tag{2.1.9}$$

Solve (2.1.9), we got the solution of optimization problem (2.1.8):

$$\hat{\theta}_{i_j}^{L_1} = \frac{\sum_{d \in S}(y_i(d) + t)x_j}{\sum_{j=1}^{v} \sum_{d \in S}(y_i(d) + t)x_j} = \frac{\sum_{d \in S}(y_i(d) + t)x_j}{m(|C_i| + t|S|)} \tag{2.1.10}$$

For estimator $\hat{\theta}_{i_j}^{L_1}$, we have the following result:

Theorem 2.1.2. *Assume for each class, we have prior distributions $p_1, p_2, \cdots, p_k$ with $p_i = |C_i|/|S|$, and we have normalized length for each document, that is: $\sum_{j=1}^{v} x_j = m$. The estimator* (2.1.10) *satisfies following property:*

1. *$\hat{\theta}_{i_j}^{L_1}$ is biased, with: $|E[\hat{\theta}_{i_j}^{L_1}] - \theta_{i_j}| = O(t)$*

2. *If* $t \leq 1$, $E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2] = O(\frac{1}{mt|S|})$. *Otherwise, the variance has the order* $O(\frac{1}{m|S|})$

Proof. 1. With assumption $\sum_{j=1}^{v} x_j = m$, we have:

$$\begin{aligned} E[\hat{\theta}_{i_j}^{L_1}] &= \frac{\sum_{d\in S}(y_i(d)+t)E[x_j]}{m(t|S|+|C_i|)} \\ &= \frac{\sum_{d\in S} tE[x_j] + \sum_{x\in C_i} E[x_j]}{m(t|S|+|C_i|)} \\ &= \frac{t\sum_{l=1}^{k}|C_l|\theta_{l_j} + \theta_{i_j}|C_i|}{t|S|+|C_i|} \\ &= \frac{t|S|\sum_{l=1}^{k} p_l\theta_{l_j} + \theta_{i_j}|C_i|}{t|S|+|C_i|} \end{aligned}$$

Thus:

$$\begin{aligned} |E[\hat{\theta}_{i_j}^{L_1}] - \theta_{i_j}| &= \frac{t|S||\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|}{t|S|+|C_i|} \\ &= \frac{|\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|}{1+p_i/t} \\ &= O(t). \end{aligned} \tag{2.1.11}$$

This shows our estimator is biased. The error is controlled by t. When t converges to 0, our estimator converges to the unbiased Naive Bayes estimator. We can also derive a lower bound for the square error:

$$\begin{aligned} E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2] &\geq (E[\hat{\theta}_{i_j}^{L_1}] - \theta_{i_j})^2 \\ &= \frac{|\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|^2}{(1+p_i/t)^2} \end{aligned}$$

2. For variance part, since

$$\hat{\theta}_{i_j}^{L_1} = \frac{\sum_{d\in S}(y_i(d)+t)x_j}{m(|C_i|+t|S|)},$$

we have:

$$
\begin{aligned}
&E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2] \\
=&E\left[\left|\frac{\sum_{d\in S}(y_i(d)+t)(x_j - E[x_j])}{m(|C_i|+t|S|)}\right|^2\right] \\
=&\frac{\sum_{d\in S}(y_i(d)+t)^2 E(|x_j - E[x_j]|^2)}{m^2(|C_i|+t|S|)^2} \\
=&\frac{\sum_{d\in C_i}(1+t)^2 m\theta_{i_j}(1-\theta_{i_j}) + \sum_{d\in C_l, l\neq i} t^2 m\theta_{l_j}(1-\theta_{l_j})}{m^2(|C_i|+t|S|)^2} \\
=&\frac{|C_i|(1+2t)\theta_{i_j}(1-\theta_{i_j}) + \sum_{l=1}^{k}|C_l|t^2\theta_{l_j}(1-\theta_{l_j})}{m(|C_i|+t|S|)^2} \\
=&\frac{|S|p_i(1+2t)\theta_{i_j}(1-\theta_{i_j}) + |S|\sum_{l=1}^{k} p_l t^2\theta_{l_j}(1-\theta_{l_j})}{m(|S|p_i+t|S|)^2} \\
=&\frac{p_i(1+2t)\theta_{i_j}(1-\theta_{i_j}) + \sum_{l=1}^{k} p_l t^2\theta_{l_j}(1-\theta_{l_j})}{m|S|(p_i+t)^2} \\
=&O(\frac{1}{m|S|})
\end{aligned}
\tag{2.1.12}
$$

□

We can see that $E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2]$ is in $O(\frac{1}{|S|})$, which means it converges faster than standard Naive Bayes $O(\frac{1}{|C_i|})$, however, since $E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|] \neq 0$, it is not an unbiased estimator.

2.1.5 Determine the correlation factor

In general statistical estimation theory, a biased estimator is acceptable, and sometimes even outperforms an unbiased estimator. A more important perspective is to find a suitable loss function to determine parameters. [60] introduces ways of choosing loss function for many famous models, the common idea is to sum a biased term and complexity penalty term for model parameters. [61] uses the maximum entropy as the loss function for text classification problems.

In our problem, from 2.1.1 and 2.1.2, we know that traditional Naive Bayes estimator is

unbiased. Our estimator is biased, but we want to find an optimal t to get smaller variance. In order to balance the trade-off between bias and variance, we would like to select a loss function which takes into account of both bias and variance.

In this task, we can use mean squared error as loss function. There is a well-known bias-variance decomposition for mean square error.

$$\begin{aligned} E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2] &= E[|\hat{\theta}_{i_j}^{L_1} - E\hat{\theta}_{i_j}^{L_1} + E\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2] \\ &= E[|\hat{\theta}_{i_j}^{L_1} - E\hat{\theta}_{i_j}^{L_1}|^2] + [E\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}]^2 \\ &:= Var(\hat{\theta}_{i_j}^{L_1}) + Bias(\hat{\theta}_{i_j}^{L_1})^2 \end{aligned}$$

In practice, we would like to minimize a general linear combination of bias and variance, namely,

$$L(\theta_{i_j}, c_1, c_2) = c_1 Bias(\hat{\theta}_{i_j}^{L_1})^2 + c_2 Var(\hat{\theta}_{i_j}^{L_1}) \tag{2.1.13}$$

Theorem 2.1.3. *The minimum of the loss function 2.1.13 is achieved at*

$$t^* = \frac{c_2(1-p_i)\theta_{i_j}(1-\theta_{i_j})}{c_2(\sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j})) - c_2\theta_{i_j}(1-\theta_{i_j}) + c_1 m|S|(\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j})^2} \tag{2.1.14}$$

Proof. First let us fix some notations for constants do not involve t to simplify the derivation. Let $\Theta_{i_j} := \theta_{i_j}(1-\theta_{i_j})$, $A = \sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j})$ and $B = (\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j})^2$. As shown in equation 2.1.11, the squared bias is

$$\begin{aligned} Bias(\hat{\theta}_{i_j}^{L_1})^2 = [E\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}]^2 &= \frac{m|S|t^2(\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j})^2}{m|S|(p_i+t)^2} \\ &= \frac{t^2 m|S|B}{m|S|(p_i+t)^2} \end{aligned}$$

From 2.1.12, the variance is

$$\begin{aligned} Var(\hat{\theta}_{i_j}^{L_1}) &= E[|\hat{\theta}_{i_j}^{L_1} - E\hat{\theta}_{i_j}^{L_1}|^2] \\ &= \frac{p_i(1+2t)\theta_{i_j}(1-\theta_{i_j}) + t^2\sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j})}{m|S|(p_i+t)^2} \\ &= \frac{p_i(1+2t)\Theta_{i_j} + t^2 A}{m|S|(p_i+t)^2} \end{aligned}$$

Therefore,

$$L(\theta_{i_j}, c_1, c_2) = \frac{c_2 p_i(1+2t)\Theta_{i_j} + t^2(c_2 A + c_1 m|S|B)}{m|S|(p_i+t)^2}$$

Then we should optimize t to minimize the loss $L(\theta_{i_j}, c_1, c_2)$. Taking derivative with respect to t and setting it to be 0. We find

$$[c_2 p_i\Theta_{i_j} + t(c_2 A + c_1 m|S|B)](p_i+t) - [c_2 p_i(1+2t)\Theta_{i_j} + t^2(c_2 A + c_1 m|S|B)] = 0$$

That simplifies to

$$c_2(p_i - 1 - t)\Theta_{i_j} + t(c_2 A + c_1 m|S|B) = 0$$

which shows

$$t = \frac{c_2(1-p_i)\Theta_{i_j}}{c_2(A - \Theta_{i_j}) + c_1 m|S|B}$$

Plug in original parameters, we obtain

$$t = \frac{c_2(1-p_i)\theta_{i_j}(1-\theta_{i_j})}{c_2\sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j}) - c_2\theta_{i_j}(1-\theta_{i_j}) + c_1 m|S|(\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j})^2}$$

□

We can see from (2.1.14) that the optimal correlation factor t^* should be a very small

number close to $O(\frac{1}{m|S|})$. Therefore by equation 2.1.11, we know the squared bias is

$$Bias(\hat{\theta}_{i_j}^{L_1})^2 = O(t^2) = O(\frac{1}{m^2|S|^2})$$

We have already shown in 2.1.12 that the order of the variance

$$Var(\hat{\theta}_{i_j}^{L_1}) = O(\frac{1}{m|S|})$$

Therefore in the case of expected square error $E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2]$ ($c_1 = c_2 = 1$) is dominated by the variance. Thus we have the following Corollary:

Theorem 2.1.4. *With any selection of $t = O(\frac{1}{m|S|})$, we have*

$$E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2] = O(\frac{1}{m|S|}) \tag{2.1.15}$$

By Theorem 2.1.1, we know that for Naive Bayes, $E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = O(\frac{1}{|C_i|})$, thus we can see that our estimator actually works better.

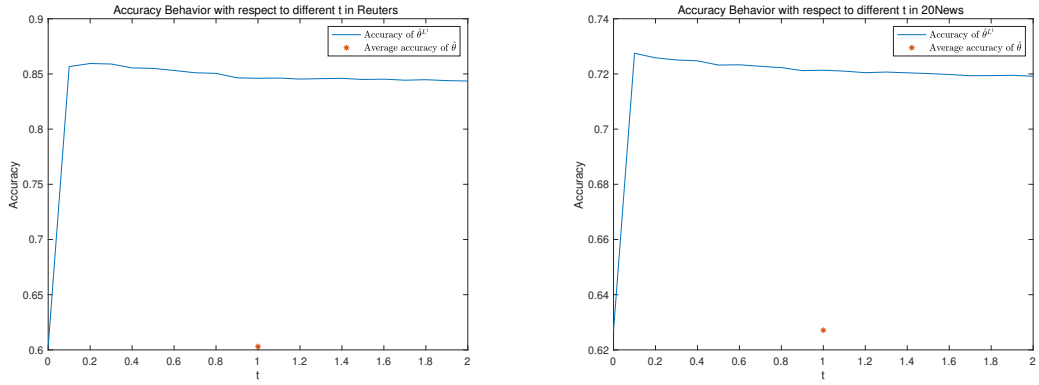
2.1.6 Experiment

Simulation with Different Correlation Factors

In the previous section we obtained that the order of t must be $O(\frac{1}{|S|})$. However, we will still need to determine how to choose the best correlation factor t. That we will have to tune the parameter by running t in some determined interval.

We applied our method on single labeled documents of 10 topics, which have almost the same sample size, in Reuters-21578 data [62], there are approximately 3000 documents in this sample set. For 20 news group data [63], it includes 20 groups and approximately 20000 documents.

We take $t \in (0, 2)$ and use 10% of data for training and assess the trained model on the remaining test data.



(a) (b)

Figure 2.1: We test accuracy behavior with respect to different correlation factors in Reuter-21578 (a) and 20 News group dataset (b). We take 10% of the data as training set. The y-axis is the accuracy and the x-axis is the correlation factor t

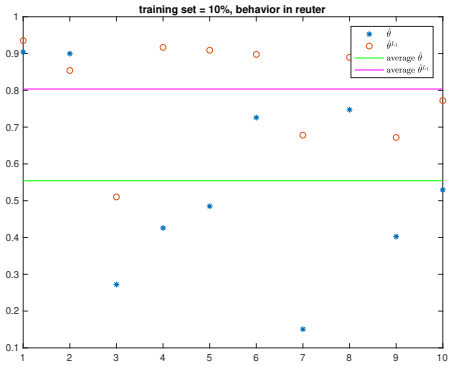
In our simulation, we notice that when we choose correlation factor to be around 0.1, we get the best accuracy for our estimation. See Figure.2.1(a) and Figure.2.1(b).

Compare with Naive Bayes

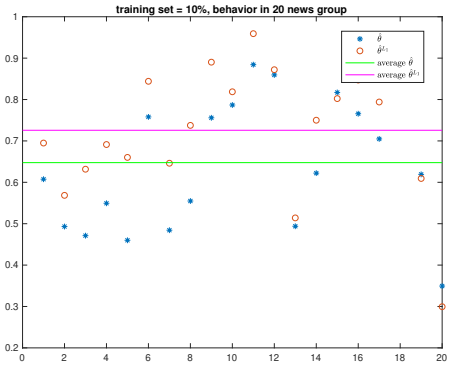
Next, we compare the result of traditional Naive Bayes estimator (2.1.4) $\hat{\theta}_{i_j}$and our estimator (2.1.10) $\hat{\theta}_{i_j}^{L_1}$. In this simulation, our correlation factor t is chosen to be 0.1 for Figure.2.2, Figure.2.3 and Figure.2.4.

First, we run both algorithms on these two sample data sets. We know that when the sample size becomes large enough, our estimator is baised. But when the training set is small, our estimator should converge faster. Thus we first take the training size relatively small (10%). See Figure.2.2(a) and Figure.2.2(b). According to the simulation, we can see our method is more accurate for most of the classes, and more accurate on average.

Then we test our estimator $\hat{\theta}^{L_1}$ with larger training set (90%). In our analysis above, we know that as datasets become large enough, our estimator converges to a biased estimator, so we expect a better result with traditional Naive Bayes estimator. See Figure.2.3(a) and Figure.2.3(b). According to the simulation, we can see for 20 news group, traditional Naive Bayes performs better than our method, but our method is still more accurate than Naive

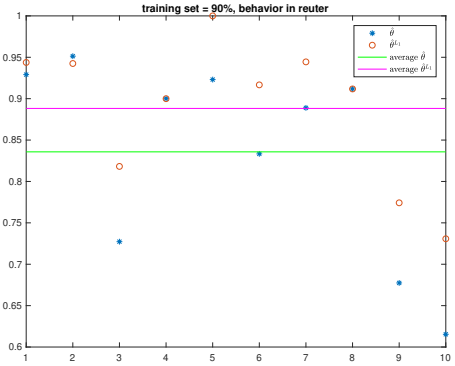


(a)

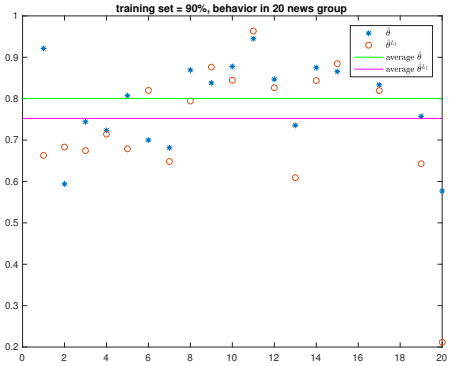


(b)

Figure 2.2: We take 10 largest groups in Reuter-21578 dataset (a) and 20 news group dataset (b), and take 10% of the data as training set. The y-axis is the accuracy, and the x-axis is the class index.



(a)

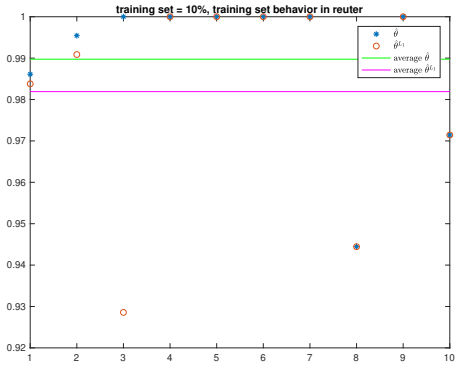


(b)

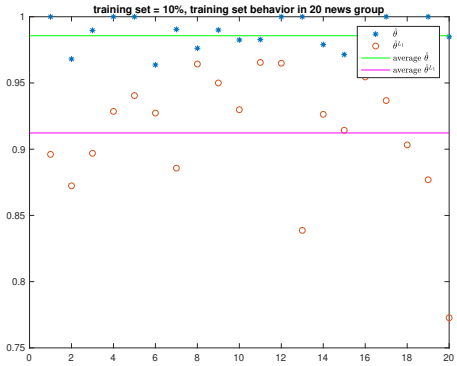
Figure 2.3: We take 10 largest groups in Reuter-21578 dataset (a) and 20 news group dataset (b), and take 90% of the data as training set. The y-axis is the accuracy, and the x-axis is the class index.

Bayes in Reuter's data. The reason might be that we have a huge unbalanced sample size in Reuter's data, 90% of the training set is still not large enough for many classes.

Finally, we apply the same training set with training size 10% and test the accuracy on the training set instead of the test set. We find the traditional Naive Bayes estimator actually achieves better results, which means it might have more over-fitting problems. This might be the reason why our method works better when the dataset is not too large: adding the correlation factor t helps us bring some uncertainty in training process, which helps avoid over-fitting. See Figure.2.4(a) and Figure.2.4(b).



(a)



(b)

Figure 2.4: We take 10 largest groups in Reuter-21578 dataset (a), and 20 news group dataset (b), and take 10% of the data as training set. We test the result on training set. The y-axis is the accuracy, and the x-axis is the class index.

Robustness of t for prediction

For estimation purposes, t must satisfy $t = O(\frac{1}{|S|})$ by theorem 2.1.4 in order to find the most accurate parameters. However, it turns out that for prediction purposes, there is a phase transition phenomenon. As long as $t \geq O(\frac{1}{|S|})$, the prediction power is not reduced even when we increase t to very large value (for example $t = 10^5$). In the simulation of finding best t in Figure.2.1(a) and Figure.2.1(b), we see the testing error is only decreasing slightly as t increasing from 0.1 to 2. We summarize this fact as follows

Proposition 2.1.1. For prediction purpose, the correlation factor t can take value in the interval

$$\frac{1}{|S|} \leq t \leq 1$$

The reason we restraint the upper bound to be 1 is that the effect of correlation factor should not exceed the effect of original class $y_i = 1$.

2.2 A Cost-Reducing Partial Labeling Estimator in Text Classification Problem

In this section, we are going to introduce the partial labeling problem, and illustrate how to apply our methods to solve it.

2.2.1 Introduction

In some circumstances, the process of labeling is distributed among less-than-expert assessors. With the fact that some data may belong to several classes by nature, their labeling for hundreds of pictures, texts, or messages a day is error-prone. The invention of partial labeling seeks to remedy the labor: instead of assigning one or some exact labels, the annotators can offer a set of possible candidate solutions for one sample, thus providing a buffer against potential mistakes [1, 4, 8, 16, 17, 26]; Other partial labeling settings involve repeated labeling to filter out noises, or assessing the quality of the labelers [18,22] to enhance the reliability of the models.

As the data size in companies such as FANG(Facebook, Amazon, Netflix, Google) constantly reaches the magnitude of Petabyte, the demand for quick, yet still precise labeling is ever growing. Viewing some practices, the partial labeling frameworks that we know exhibit some limitations. For instance, in a real-world situation concerning NLP, if the task is to determine the class/classes of one article, an annotator with a bachelor degree of American literature might find it difficult to determine if an article with words dotted with 'viscosity', 'gradient', and 'Laplacian' etc. belongs to computer science, math, physics, chemistry, or none of the classes above. As a result, the annotator might struggle within some limited amount of time amid a large pool of label classes and is likely to make imprecise choices even in a lenient, positive-oriented partial labeling environment. Another issue is the cost. Repeated labeling and keeping track of the performance of each labeler may be pricey, and the anonymity of the labelers can raise another barrier wall to certain partial labeling approaches.

Taking the real world scenarios into consideration, we present a new method to tackle the problem on how to gather at a large scale partially correct information from diverse annotators, while remaining efficient and budget-friendly. Still taking the above text classification problem as the example. Although that same annotator might not easily distinguish which categories the above-mentioned article belongs to, in a few seconds he/she can rule out the possibility the article is related to cuisines, TV-entertainment, or based on his/her own expertise, novels. In our partial labeling formulation, the safe choices, crossed-off categories labeled by annotators can still be of benefit. Furthermore, when contradictory labels are marked on one training sample and the identities of the labelers unknown, our introduced self-correcting estimator can select, and learn from the categories where the labels agree.

Based on this, our research proposes a new way to formulate partial labeling. For some documents, instead of having exact labels, not belonging to certain classes is the information provided, which we will take as negative labeling. To make use of both kinds of data, we propose two maximum likelihood estimators, one of which has a self-correction property to estimate the distribution of each classes. By making both type of labeled data contribute in the training process, we prove that new estimators converge faster than traditional Naive Bayes estimator. Finally we find a way to apply new methods to some only positively labeled data set, which is identified as a fully supervised learning problem, and achieve a better result compared to the traditional Naive Bayes.

2.2.2 Related work

The text classification problem is seeking a way to best distinguish different types of documents[64, 65]. Being a traditional natural language processing problem, one needs to make full use of the words and sentences, converting them into various input features, and applying different models to process training and testing. A common way to convert words into features is to encoding them based on the term frequency and inverse document frequency,

as well as the sequence of the words. There are many results about this, for example, tf-idf[66] encodes term t in document d of corpus D as:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D),$$

where $tf(t, d)$ is defined as term frequency, it can be computed as $tf(t, d) = \frac{|t:t \in d|}{|d|}$, and $idf(t, D)$ is defined as inverse document frequency, it can be computed as

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}.$$

We also have n-gram techniques, which first combines n nearest words together as a single term, and then encodes it with tf-idf. Recently, instead of using tf-idf, [67] defines a new feature selection score for text classification based on the KL-divergence between the distribution of words in training documents and their classes.

A popular model to achieve our aim is to use Naive Bayes model[40, 41], the label for a given document d is given by:

$$label(d) = \operatorname*{argmax}_j P(C_j)P(d|C_j),$$

where C_j is the j-th class. For example, we can treat each class as a multinomial distribution, and the corresponding documents are samples generated by the distribution. With this assumption, we desire to find the centroid for every class, by either using the maximum likelihood function or defining other different objective functions[54] in both supervised and unsupervised learning version[49]. Although the assumption of this method is not exact in this task, Naive Bayes achieves high accuracy in practical problems.

There are also other approaches to this problem, one of which is simply finding linear boundaries of classes with support vector machine[45, 44]. Recurrent Neural Network (RNN)[46, 47] combined with word embedding is also a widely used model for this prob-

lem.

In real life, there have been many researches on this problem. Most of researches focus on the situation where there are a set of candidate labels available for the training data instead of an exact label. Some researches mainly focus on the learning strategy with some novel loss functions. One may have different type of labels[68], in which circumstance, semi-supervised learning or partial-label problems need to be considered [69]. There are several methods to encode the partial label information into the learning framework. For the partial label data set, one can define a new loss combining all information of the possible labels, for example, in [70], the authors introduce a discriminative learning approach that incorporates partial label information into the conventional margin-based learning framework and modify the traditional L^2 loss:

$$L(w) = \frac{1}{n+m}\left[\sum_{i=1}^{n} l(x_i, y_i, w) + \sum_{i=1}^{m} l(x_i, Y_i, w)\right],$$

where Y_i is the possible label set for x_i and $l(x_i, Y_i, w)$ is a non-negative loss function. In order to utilizing the same L-2 norm regularization and the hinge loss for fully labeled data, authors propose the loss function $l(x_i, Y_i, w)$ for partial label data as:

$$\max(0, 1 - [\max_{y_i \in Y_i} w^T \Phi(x_i.y_i) - \max_{\bar{y_i} \notin Y_i} w^T \Phi(x_i, \bar{y_i})])$$

In [69], authors propose a type of one-vs-all scheme for the supervised case and define the convex loss for partial labels as:

$$L_\Psi(g(x), y) = \Psi(\frac{1}{|y|}\sum_{a \in y} g_a(x)) + \sum_{a \notin y} \Psi(-g_a(x))$$

where Ψ is convex, differentiable and bounded below, for which all of exponential, logistic and squared hinge loss are satisfied. Then y is a singleton and $g_a(x)$ is a score function for label a as input x.

A modification of the likelihood function is as well an approach to this problem and [71] gives the following optimization problem using Naive Bayes method

$$\theta^* = \arg\max_{\theta} \sum_{i} \sum_{y_i \in S_i} p(y|x_i, \theta)$$

where S_i is the possible labels for x_i.

In [72], authors propose a n instance-based approach, which directly disambiguating the candidate label set using an iterative label propagation procedure. Then the model will do classification by minimizing error reconstruction from its nearest neighbors with the following loss function:

$$\min_{\omega_j} ||x_j - \sum_{a=1}^{k} \omega_{i_a,j} x_{i_a}||^2$$

In [73], authors propose the idea of self-training, which utilize a unified formulation to train the model and perform pseudo-labeling jointly. For pseudo-labeling, authors introduce the maximum infinity norm regularization on the modeling outputs, which leads to a convex-concave optimization problem. The optimization problem is as follows:

$$\min_{P,W,b} \sum_{i=1}^{m} (||W^T x_i + b - p_i||_2^2 - \lambda ||p_i||_\infty) + \beta ||W||_F^2$$

A semi-supervised partial label learning method introduced in [74], which is an iterative label propagation procedure between partial labeled data and unlabeled data to disambiguate the candidate label sets of partial labeled samples.

There are also some researches focusing on feature space. In [75], authors propose a new learning strategy via label enhancement. The idea is to recover the generalized label distribution by using the topological information of the feature space, which is different from other methods that focus on disambiguating the candidate label sets. The optimization

problem for recovery is defined as follows:

$$\min_{\hat{W}} tr[(\hat{W}\Phi - L)^T(\hat{W}\Phi - L)] + \lambda tr(\hat{W}\Phi G\Phi^T\hat{W}^T)$$

And in [76], authors propose a two-stage approach based on feature-aware disambiguation. To generate normalized label confidence, authors introduce the optimization problem as:

$$min \sum_{i=1}^{m}\sum_{j=1}^{m} \gamma_{ij}\lambda_i^T\lambda_j$$

and then they induce the regression model to finish classification by minimizing the following loss function:

$$L(\Theta, b) = \frac{1}{2}\sum_{k=1}^{q}||\theta_k||^2 + C_1\sum_{i=1}^{m}L_1(u_i) + C_2\sum_{i=1}^{m}v_i$$

Meanwhile, the similarity of features among data could be considered to give a confidence of each potential labels for a certain data. In [76], K nearest neighbor (KNN) is adopted to construct a graph structure with the information of features while Rocchio and Rocchio with clustering are used in [68].

2.2.3 General Setting

Different from other researches, we focus on the situation where we don't need a candidate label set for partial labeled data. Instead, we will have a data set, which contains negative labeled data. Therefore, we will consider a classification problem with sample $x \in S$ and class set C, where

$$C = \{C_1, C_2, ..., C_k\}.$$

We are interested in finding our estimator:

$$\hat{y} = f(x;\theta) = (f_1(x;\theta), f_2(x;\theta), ..., f_k(x;\theta))$$

for y, where $\theta = \{\theta_1, \theta_2, ..., \theta_m\}$ is the parameter, and $f_i(x; \theta)$ is the likelihood function of sample x in class C_i.

Then without further notification:

- assume that all the categories are independent multinomial distributions and each document is a sample independently generated by a certain distribution
- S is the document set such that $S = S_1 \cup S_2$ and assume the class set C has k different categories: $\{C_1, C_2, ..., C_k\}$.
- dataset S_1: we know exactly that sample d is in a class, and not in other classes. In this case, define: $y = (y_1, y_2, ..., y_k)$, if d is in class C_i, then $y_i = 1$. Notice that if this is a single label problem, then we have: $\sum_{i=1}^{k} y_i = 1$
- dataset S_2: we only have the information that sample d is not in a class, then $y_i = 0$. In this case, define: $z = (z_1, z_2, ..., z_k)$, if d is not in class C_i, we have $z_i = 1$
- for each category C_i, the centroid $\theta_i = (\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_v})$ and θ_i satisfies: $\sum_{j=1}^{v} \theta_{i_j} = 1$
- Assume we have totally v different words, thus for each document $d \in S$:$d = \{x_1, x_2, \cdots, x_v\}$. ,where x_i represents the number of occurrence for i-th word with $\sum_{j=1}^{v} x_j = m$
- Assume label vector $y = (y_1, y_2, \cdots, y_k)$. For document d in class C_i, $y_i(d) = 1$ and $\sum_{i=1}^{k} y_i = 1$.
- $\hat{y}(d) = f(d; \theta) = (f_1(d; \theta), f_2(d; \theta), ..., f_k(d; \theta))$ is our estimator for y, where θ is the parameter matrix and $f_i(d; \theta)$ is the likelihood function of document d in class C_i.

To build the model, we define the following likelihood ratio function and likelihood

function:

$$L_1(\theta) = \prod_{x\in S_1}\prod_{i=1}^{k} f_i(x;\theta)^{y_i} \prod_{x\in S_2}\prod_{i=1}^{k} f_i(x;\theta)^{\frac{1-z_i}{k-\sum_{j\neq i} z_j}}. \tag{2.2.1}$$

$$L_2(\theta) = \prod_{x\in S} \frac{\prod_{i=1}^{k} f_i(x;\theta)^{y_i(x)+t}}{\prod_{i=1}^{k} f_i(x;\theta)^{z_i(x)}} = \prod_{x\in S}\prod_{i=1}^{k} f_i(x;\theta)^{y_i(x)-z_i(x)+t}. \tag{2.2.2}$$

The t in L_2 satisfy $t > 1$, which is a parameter to avoid non-convexity.

The intuition of L_1 is to consider the sample labeled $z_i = 1$ has equal probability to be labeled in the other classes, each of the classes will have probability $\frac{1-z_i}{k-\sum_{j\neq i} z_j}$. And the intuition of L_2 is to consider this in a likelihood ratio way, the $z_i = 1$ labeled sample will have negative affection for class C_i, so we put it in the denominator. With $t > 1$, all the terms in denominator will finally be canceled out, so that even $f_i(x;\theta) = 0$ for some sample $x \in S$ will not cause trouble. Another intuition for L_2 is that, it can be self-correct the repeated data, which has been labeled incorrectly.

Take logarithm for both side, we obtain the following functions:

$$\log(L_1(\theta)) = \sum_{x\in S_1}\sum_{i=1}^{k} y_i(x)\log f_i(x,\theta) + \sum_{x\in S_2}\sum_{i=1}^{k} \frac{1-z_i}{k-\sum_{j\neq i} z_j}\log f_i(x,\theta), \tag{2.2.3}$$

and

$$\log(L_2(\theta)) = \sum_{x\in S}\sum_{i=1}^{k} (y_i(x) + t - z_i(x))\log f_i(x,\theta). \tag{2.2.4}$$

We would like to find our estimator $\hat{\theta}$ such that (2.2.4) or (2.2.3) reaches maximum.

2.2.4 Main Result

From Theorem.2.1.1, we can see that traditional Naive Bayes estimator $\hat{\theta}$ is an unbiased estimator with variance $O(\frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m})$. Now we are trying to solve our estimators, and prove they can use the data in dataset S_2, and perform better than traditional Naive Bayes estimator.

Text classification with L_1 setting (2.2.1)

In order to use data both in S_1 and S_2, we would like to solve (2.1.3) with $L(\theta) = L_1(\theta)$, where L_1 is defined as (2.2.1), let:

$$G_i = 1 - \sum_{j=1}^{v} \theta_{i_j},$$

by Lagrange multiplier, we have:

$$\begin{cases} \dfrac{\partial \log(L_1)}{\partial \theta_{i_j}} + \lambda_i \dfrac{\partial G_i}{\partial \theta_{i_j}} = 0 \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases}$$

Plug in, we obtain:

$$\begin{cases} \displaystyle\sum_{x \in S_1} \frac{y_i(x) x_j}{\theta_{i_j}} + \sum_{x \in S_2} \frac{1 - z_i(x)}{k - \sum_{l \neq i} z_l(x)} \cdot \frac{x_j}{\theta_{i_j}} - \lambda_i = 0, \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases} \tag{2.2.5}$$

Solve (2.2.5), we got the solution of optimization problem (2.1.3):

$$\hat{\theta}_{i_j}^{L_1} = \frac{\sum_{x \in S_1} y_i(x) x_j + \sum_{x \in S_2} \frac{1 - z_i(x)}{k - \sum_{l \neq i} z_l(x)} x_j}{\sum_{x \in S_1} y_i(x) \sum_{j=1}^{v} x_j + \sum_{x \in S_2} \frac{1 - z_i(x)}{k - \sum_{l \neq i} z_l(x)} \sum_{j=1}^{v} x_j}. \tag{2.2.6}$$

Theorem 2.2.1. *Assume we have normalized length of each document, that is: $\sum_{j=1}^{v} x_j = m$ for all d. Let $Z_i(x) = \frac{1 - z_i(x)}{k - \sum_{l \neq i} z_l(x)} = K$, $l_{i_j} = E[x_j | Z_i = K]/m$. Assume further that $|\{i : z_i(x) = 1\}| = K$ to be a constant for all $x \in S_2$, the estimator* (2.2.6) *satisfies following properties:*

1. *$\hat{\theta}_{i_j}^{L_1}$ is biased with*

$$E[\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}] = \frac{|R_i| K (l_{i_j} - \theta_{i_j})}{|C_i| + |R_i| K}.$$

2. $E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2] = O\left(\frac{1}{|S_1|+|S_2|}\right)$.

Proof. 1. We denote $Z_i(x) = \frac{1-z_i(x)}{k-\sum_{l\neq i} z_l(x)} = K$, $l_{i_j} = E[x_j|Z_i = K]/m$ and $R_i = \{x : z_i(x) = 0\}$, we have:

$$\hat{\theta}_{i_j}^{L_1} = \frac{\sum_{x\in S_1} y_i(x)x_j + \sum_{x\in S_2} Z_i(x)x_j}{(\sum_{x\in S_1} y_i(x) + \sum_{x\in S_2} Z_i(x))m} = \frac{\sum_{x\in S_1} y_i(x)x_j + \sum_{x\in S_2} Z_i(x)x_j}{(|C_i| + |R_i|K)m}$$

Moreover, assuming that $p_i = P(y_i(x) = 1) = |C_i|/|S_1|, q_i = P(z_i(x) = 0) = |R_i|/|S_2|$, it holds that

$$\begin{aligned}
E[\hat{\theta}_{i_j}^{L_1}] &= \frac{1}{(|C_i| + |R_i|K)m}\left(\sum_{x\in S_1} E[y_i(x)x_j] + \sum_{x\in S_2} E[Z_i(x)x_j]\right) \\
&= \frac{1}{(|C_i| + |R_i|K)m}\sum_{x\in S_1} p_i E[x_j|y_i(x) = 1] \\
&+ \frac{1}{(|C_i| + |R_i|K)m}\sum_{x\in S_2} q_i K E[x_j|Z_i(x) = K] \\
&= \frac{|C_i|E[x_j|y_i(x) = 1] + |R_i|KE[x_j|Z_i(x) = K]}{(|C_i| + |R_i|K)m} \\
&= \frac{|C_i|\theta_{i_j}m + |R_i|Kl_{i_j}m}{(|C_i| + |R_i|K)m}.
\end{aligned}$$

Thus,

$$E[\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}] = \frac{|R_i|K(l_{i_j} - \theta_{i_j})}{|C_i| + |R_i|K}$$

2. As is for the second part, we have

$$\begin{aligned}
\left(\hat{\theta}_{i_j}^{L_1}\right)^2 = \frac{1}{(|C_i| + |R_i|K)m}\Big(&\sum_{\alpha\in S_1}\sum_{\beta\in S_1} y_i(\alpha)y_i(\beta)\alpha_j\beta_j \\
&\sum_{\alpha\in S_2}\sum_{\beta\in S_2} Z_i(\alpha)Z_i(\beta)\alpha_j\beta_j \\
&\sum_{\alpha\in S_1}\sum_{\beta\in S_2} y_i(\alpha)Z_i(\beta)\alpha_j\beta_j\Big).
\end{aligned}$$

Then, by introducing $C = (|C_i| + |R_i|K)m$ and $L_{i_j} = E[x_j^2|Z_i(x) = K]$ it is true

that

$$
\begin{aligned}
E\left[\left(\hat{\theta}_{i_j}^{L_1}\right)^2\right] =& \frac{1}{C^2}\Big(\quad \sum_{x\in S_1} E[y_i^2(x)x_j^2] + \sum_{\alpha,\beta\in S_1,\alpha\neq\beta} E[y_i(\alpha)\alpha_j]E[y_i(\beta)\beta_j] \\
&+ \sum_{x\in S_2} E[Z_i^2(x)x_j^2] + \sum_{\alpha,\beta\in S_2,\alpha\neq\beta} E[Z_i(\alpha)\alpha_j]E[Z_i(\beta)\beta_j] \\
&+ 2\sum_{\alpha\in S_1,\beta\in S_2} E[y_i(\alpha)\alpha_j]E[Z_i(\beta)\beta_j]\Big) \\
=& \frac{1}{C^2}\Big(\quad |C_i|m\theta_{i_j}(1-\theta_{i_j}+m\theta_{i_j}) + \left(|S_1|^2-|S_1|\right)p_i^2m^2\theta_{i_j}^2 \\
&+ |R_i|K^2L_{i_j} + \left(|S_2|^2-|S_2|\right)K^2q_i^2m^2l_{i_j}^2 \\
&+ 2|C_i||R_i|m^2K\theta_{i_j}l_{i_j}|\Big) \\
=& \frac{1}{C^2}\Big(\quad |C_i|m\theta_{i_j}(1-\theta_{i_j}+m\theta_{i_j}) - |S_1|p_i^2m^2\theta_{i_j}^2 \\
&+ |R_i|K^2L_{i_j} - |S_2|K^2q_i^2m^2l_{i_j}^2\Big) \\
&+ \left(\frac{|C_i|\theta_{i_j}m + |R_i|Kl_{i_j}m}{(|C_i|+|R_i|K)m}\right)^2
\end{aligned}
$$

Using the fact that $E\left|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]\right|^2 = E\left[\left(\hat{\theta}_{i_j}^{L_1}\right)^2\right] - \left(E[\hat{\theta}_{i_j}^{L_1}]\right)^2$, we can conclude that

$$
\begin{aligned}
&E\left|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]\right|^2 \\
&= \frac{1}{(|C_i|+|R_i|K)^2m^2} \qquad \Big(|C_i|m\theta_{i_j}(1-\theta_{i_j}+m\theta_{i_j}) - |S_1|p_i^2m^2\theta_{i_j}^2 \\
&\qquad\qquad\qquad\qquad + |R_i|K^2L_{i_j} - |S_2|K^2q_i^2m^2l_{i_j}^2\Big) \\
&= O\left(\frac{1}{|S_1|+|S_2|}\right)
\end{aligned}
$$

□

Comparing $\hat{\theta}_{i_j}$ and $\hat{\theta}_{i_j}^{L_1}$, we can see that even though our estimator is biased, the variance of $\hat{\theta}_{i_j}^{L_1}$ is significant smaller than the variance of $\hat{\theta}_{i_j}$, which means by using negative sample set, $\hat{\theta}_{i_j}^{L_1}$ converges way faster than original Naive Bayes estimator $\hat{\theta}_{i_j}$.

Text classification with L_2 setting (2.2.2)

Another way to use both S_1 and S_2 dataset is to solve (2.1.3) with $L(\theta) = L_2(\theta)$, where L_2 is defined as (2.2.2), let:

$$G_i = 1 - \sum_{j=1}^{v} \theta_{i_j},$$

by Lagrange multiplier, we have:

$$\begin{cases} \dfrac{\partial \log(L_2)}{\partial \theta_{i_j}} + \lambda_i \dfrac{\partial G_i}{\partial \theta_{i_j}} = 0 \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases}$$

Plug in, we obtain:

$$\begin{cases} \displaystyle\sum_{x \in S} (y_i(x) + t - z_i(x)) \frac{x_j}{\theta_{i_j}} - \lambda_i = 0 \; \forall \; 1 \leq i \leq k \text{ and } \forall \; 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \; \forall \; 1 \leq i \leq k \end{cases} \tag{2.2.7}$$

Solve (2.2.7), we got the solution of optimization problem (2.1.3):

$$\hat{\theta}_{i_j}^{L_2} = \frac{\sum_{x \in S} (y_i(x) + t - z_i(x)) x_j}{\sum_{j=1}^{v} \sum_{x \in S} (y_i(x) + t - z_i(x)) x_j}. \tag{2.2.8}$$

Notice that the parameter t here is used to avoid non-convexity, when $0 \leq t < 1$, the optimization problem (2.1.3) has the optimizer located on the boundary of θ, which cannot be solved explicitly.

Theorem 2.2.2. *Assume we have normalized length of each document, that is: $\sum_{j=1}^{v} x_j = m$ for all d. Assume the negative label has only one entry to be 1, namely $\sum_i z_i(x) = 1, \forall x \in S_2$. Let $|C_i|$ denote the number of documents in Class i and $|D_i|$ denote the number of documents labelled not in Class i with $p_i = \frac{|C_i|}{|S|}$ and $q_i = \frac{|D_i|}{|S|}$. Further, we assume if a document x is labelled not in Class i, it will have equal probability to be in any other class.*

Then the estimator (2.2.8) *satisfies following properties:*

1. $\hat{\theta}_{i_j}^{L_2}$ *is biased and* $|E[\hat{\theta}_{i_j}^{L_2} - \theta_{i_j}]| = O(\frac{t+q_i}{t+p_i-q_i})$
2. $var[\hat{\theta}_{i_j}^{L_2}] = O\left(\frac{(1+2t)p_i+(1-2t)q_i+t^2}{m(p_i-q_i+t)^2|S|}\right)$.

Proof. First of all, we can simplify (2.2.8) using our assumption to be

$$\hat{\theta}_{i_j}^{L_2} = \frac{\sum_{x\in S}(y_i(x) + t - z_i(x))x_j}{\sum_{x\in S}(y_i(x) + t - z_i(x))m}.$$

For $x \in C_l \subset S_1$, $E[x_j] = m\theta_{l_j}$ and $var[x_j] = m\theta_{l_j}(1-\theta_{l_j})$. For $x \in D_l \subset S_2$ with $z_l(x) = 1$, which means x is labelled not in Class l, we have $E[x_j] = \frac{m\sum_{r\neq l}\theta_{r_j}}{k-1}$ and

$$\begin{aligned} var(x_j, x \in D_l) &= \sum_{r\neq l} E[E[(x_j - E[x_j])^2 | x \in C_r]] \\ &= \sum_{r\neq l} \frac{1}{k-1} E[(x_j - E[x_j])^2 | x \in C_r] \\ &= \frac{1}{k-1}\sum_{r\neq l} m\theta_{r_j}(1-\theta_{r_j}) \end{aligned}$$

Moreover, denote $N = m\sum_{x\in S}(y_i(x) + t - z_i(x)) = m(|C_i| - |D_i| + t|S|)$.

1. We first compute the expectation

$$\begin{aligned} & E[\hat{\theta}_{i_j}^{L_2}] \\ =& \frac{\sum_{x\in S}(y_i(x) + t - z_i(x))E[x_j]}{N} \\ =& \frac{t\sum_{x\in S_1} E[x_j] + t\sum_{x\in S_2} E[x_j] + m|C_i|\theta_{i_j} - m|D_i|\frac{\sum_{l\neq i}\theta_{l_j}}{k-1}}{N} \\ =& \frac{t\sum_{l=1}^k |C_l|\theta_{l_j} + |C_i|\theta_{i_j} + t\sum_{l=1}^k |D_l|\frac{\sum_{r\neq l}\theta_{r_j}}{k-1} - |D_i|\frac{\sum_{l\neq i}\theta_{l_j}}{k-1}}{|C_i| - |D_i| + t|S|} \\ =& \frac{t\sum_{l=1}^k p_l\theta_{l_j} + p_i\theta_{i_j} + t\sum_{l=1}^k q_l\frac{\sum_{r\neq l}\theta_{r_j}}{k-1} - q_i\frac{\sum_{l\neq i}\theta_{l_j}}{k-1}}{p_i - q_i + t} \end{aligned}$$

Therefore, we can compute the bias:

$$
\begin{aligned}
E[\hat{\theta}_{i_j}^{L_2} - \theta_{i_j}] &= E[\hat{\theta}_{i_j}^{L_2}] - \frac{(p_i - q_i + t)\theta_{i_j}}{p_i - q_i + t} \\
&= \frac{t\sum_{l=1}^{k} p_l\theta_{l_j} + t\sum_{l=1}^{k} q_l\frac{\sum_{r\neq l}\theta_{r_j}}{k-1} - t\theta_{i_j} - q_i(\frac{\sum_{l\neq i}\theta_{l_j}}{k-1} - \theta_{i_j})}{p_i - q_i + t} \\
&= O(\frac{t + q_i}{t + p_i - q_i})
\end{aligned} \tag{2.2.9}
$$

2. We now turn to variance.

$$
\begin{aligned}
var(\hat{\theta}_{i_j}^{L_2}) &= E[|\hat{\theta}_{i_j}^{L_2} - E[\hat{\theta}_{i_j}^{L_2}]|^2] \\
&= \frac{1}{N^2}E\left(\sum_{x\in S}(y_i(x) + t - z_i(x))(x_j - E[x_j])\right)^2
\end{aligned}
$$

Since different document $x \in S$ are independent, we have

$$
var(\hat{\theta}_{i_j}^{L_2}) = \frac{1}{N^2}\sum_{x\in S}(y_i(x) + t - z_i(x))^2 var(x_j)
$$

where $var(x_j) = E\left(x_j - E[x_j]\right)^2$.

If $x \in C_l$ has positive labels, $var(x_j) = m\theta_{l_j}(1 - \theta_{l_j})$. Then

$$
\begin{aligned}
V_1 :=& \sum_{x\in S_1}(y_i(x) + t - z_i(x))^2 var(x_j) \\
=& \sum_{x\in C_i}(1+t)^2 m\theta_{i_j}(1-\theta_{i_j}) + \sum_{x\in C_l, l\neq i} t^2 m\theta_{l_j}(1-\theta_{l_j}) \\
=& |C_i|(1+2t)m\theta_{i_j}(1-\theta_{i_j}) + \sum_{l=1}^{k}|C_l|t^2 m\theta_{l_j}(1-\theta_{l_j}) \\
=& O(|C_i|(1+2t)m) + O(|S_1|t^2 m)
\end{aligned}
$$

On the other hand if $x \in D_l \subset S_2$ has negative labels, $var(x_j, x \in D_l) = \frac{1}{k-1}\sum_{r\neq l} m\theta_{r_j}(1-$

$\theta_{r_j})$

$$
\begin{aligned}
V_2 :=& \sum_{x\in S_2}(y_i(x)+t-z_i(x))^2 var(x_j)\\
=& \sum_{x\in D_i}(t-1)^2 var(x_j, x\in D_i) + \sum_{l\neq i}^{k}\sum_{x\in D_l} t^2 var(x_j, x\in D_l)\\
=& \sum_{x\in D_i}(1-2t) var(x_j, x\in D_i) + \sum_{l=1}^{k}\sum_{x\in D_l} t^2 var(x_j, x\in D_l)\\
=& \frac{(1-2t)|D_i|}{k-1}\sum_{r\neq i} m\theta_{r_j}(1-\theta_{r_j}) + \frac{t^2}{k-1}\sum_{l=1}^{k}|D_l|\sum_{r\neq l} m\theta_{r_j}(1-\theta_{r_j})\\
=& O((1-2t)m|D_i|) + O(t^2 m|S_2|)
\end{aligned}
$$

Then the variance is

$$
\begin{aligned}
var(\hat{\theta}_{i_j}^{L_2}) =& \frac{1}{N^2}(V_1+V_2) = \frac{V_1+V_2}{m^2\left(|C_i|-|D_i|+t|S|\right)^2}\\
=& O(\frac{V_1+V_2}{m^2(p_i-q_i+t)^2|S|^2})\\
=& O\left(\frac{[(1+2t)p_i + \frac{|S_1|}{|S|}t^2 + (1-2t)q_i + \frac{|S_2|}{|S|}t^2]}{m(p_i-q_i+t)^2|S|}\right)\\
=& O\left(\frac{(1+2t)p_i + (1-2t)q_i + t^2}{m(p_i-q_i+t)^2|S|}\right)
\end{aligned}
$$

□

Using the same strategy as in 1, we have the first part of our variance estimation should be of order $O(\frac{1}{m|S|})$, which is less than the order of variance for Naive Bayes estimation:$O(\frac{1}{|C_i|})$. We also showed that its order is $O(\frac{1}{|S_1|+|S_2|}) < O(\frac{1}{|C_i|})$, therefore, $\hat{\theta}_{i_j}^{L_2}$ converges faster than $\hat{\theta}_{i_j}$.

Improvement of Naive Bayes estimator with only S_1 dataset

Now assume that we don't have dataset S_2, but only have dataset $S = S_1$, can we still do better than traditional Naive Bayes estimator $\hat{\theta}$? To improve the estimator, we can try to use L_1 or L_2 setting. With $z(x) = 1 - y(x)$, we can define function z on S_1 dataset.

With simple computation, we have the estimator of L_1 is the same as $\hat{\theta}_{i_j}$. as for the estimator for L_2, we have:

$$\hat{\theta}^*_{i_j} = \frac{\sum_{x \in S}(2y_i(x) + t - 1)x_j}{\sum_{j=1}^{v}\sum_{x \in S}(2y_i(x) + t - 1)x_j}, \tag{2.2.10}$$

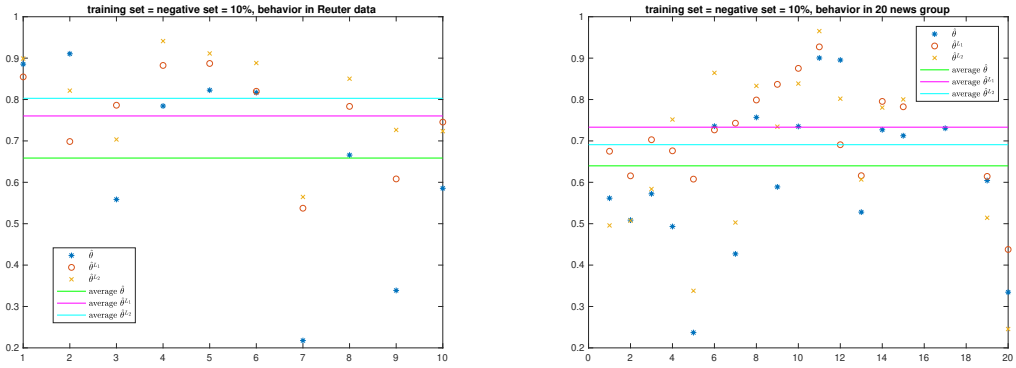
Corollary 2.2.3. *Assume we have normalized length of each document, that is: $\sum_{j=1}^{v} x_j = m$ for all d. With only dataset S_1, let $S_2 = S_1$, define $z(x) = 1 - y(x)$, Then the estimator* (2.2.10) *satisfies following properties:*

1. *$\hat{\theta}^*_{i_j}$ is biased, $E[\hat{\theta}^*_{i_j} - \theta_{i_j}] = O(t)$.*
2. *$E[|\hat{\theta}^*_{i_j} - \theta_{i_j}|^2] = O(\frac{1}{|S|})$.*

2.2.5 Experiment

We applied our method on top 10 topics of single labeled documents in Reuters-21578 data[62], and 20 news group data[63]. we compare the result of traditional Naive Bayes estimator $\hat{\theta}_{i_j}$ and our estimator $\hat{\theta}^{L_1}_{i_j}$, $\hat{\theta}^{L_2}_{i_j}$, as well as $\hat{\theta}^*_{i_j}$. t is chosen to be 2 in all the following figures. The data in S_2 is generated randomly by not belong to a class, for example, if we know a document d is in class 1 among 10 classes in Reuter's data, to put d in S_2, we randomly pick one class from 2 to 10, and mark d not in that class.

First of all, we run all the algorithms on these two sample sets. We know that when sample size becomes large enough, our estimators actually convergence into something else, but when sample size small enough, our estimator should converge faster. Thus we take the training size relatively small. See Figure.2.5(a) and Figure.2.5(b). According from



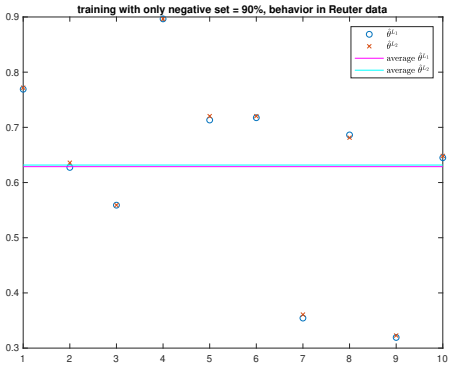
(a) (b)

Figure 2.5: We take 10 largest groups in Reuter-21578 dataset (a) and 20 news group dataset (b), and take 20% of the data as training set, among which $|S_1| = |S_2|$. The y-axis is the accuracy, and the x-axis is the class index.

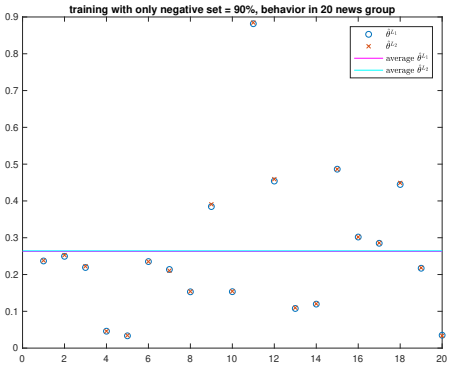
the experiment, we can see our methods are more accurate for most of the classes, and more accurate in average.

Then we consider a more extreme case. If we have a dataset with $|S_1| = 0$, that is to say, we have no positive labeled data. In this setting, traditional Naive Bayes will not work, but what will we get from our estimators? See Figure.2.6(a) and Figure.2.6(b). We can see we can still get some information from negative labeled data. The accuracy is not as good as Figure.2.5(b) and Figure.2.5(a), that is because for each of the sample, negative label is only a part of information of positive label.

At last, we test our estimator $\hat{\theta}^{L_2}$ with only S_1 dataset, see Figure.2.7(a) and Figure.2.7(b). We can see our method achieve better result than traditional Naive Bayes estimator. We try to apply same training set and test the accuracy just on training set, we find traditional Naive Bayes estimator actually achieve better result, that means it might have more over-fitting problems, see Figure.2.8(a) and Figure.2.8(b).

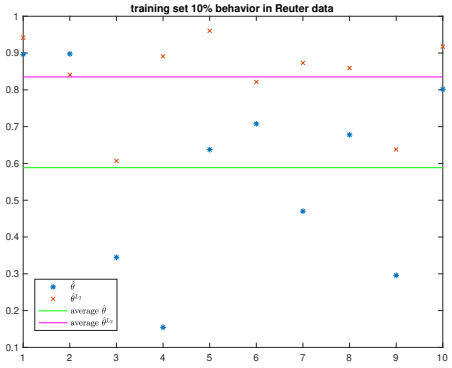


(a)

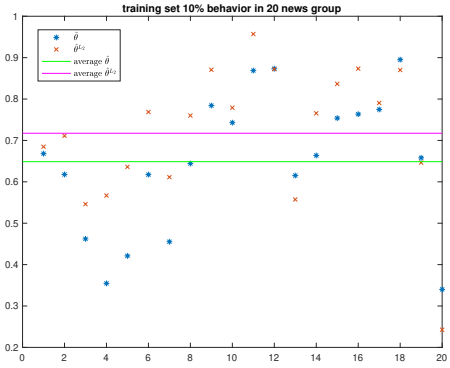


(b)

Figure 2.6: We take 10 largest groups in Reuter-21578 dataset (a), and 20 news group dataset (b), and take 90% of the data as S_2 training set. The y-axis is the accuracy, and the x-axis is the class index.

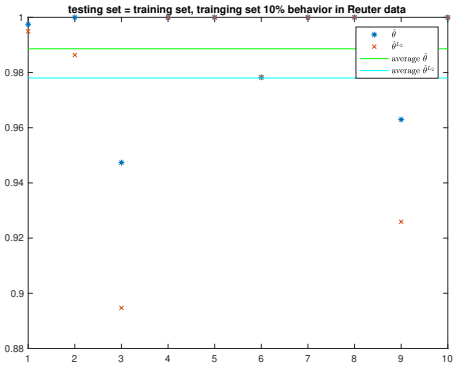


(a)

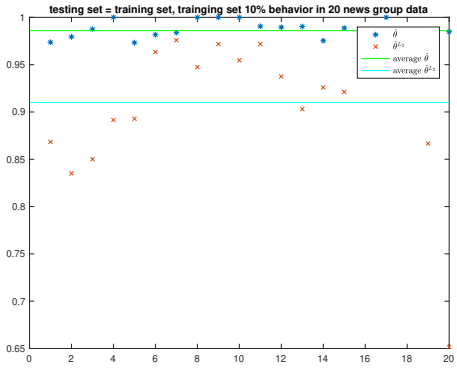


(b)

Figure 2.7: We take 10 largest groups in Reuter-21578 dataset (a), and 20 news group dataset (b), and take 10% of the data as S_1 training set. The y-axis is the accuracy, and the x-axis is the class index.



(a)



(b)

Figure 2.8: We take 10 largest groups in Reuter-21578 dataset(a), and 20 news group dataset (b), and take 10% of the data as S_1 training set. We test the result on training set. The y-axis is the accuracy, and the x-axis is the class index.

REFERENCES

[1] S. Amsalu, J. Duan, H. Matzinger, and I. Popescu, "Recovery of spectrum from estimated covariance matrices and statistical kernels for machine learning and big data," *arXiv preprint arXiv:1804.09472*, 2018.

[2] V. Koltchinskii, K. Lounici, *et al.*, "Normal approximation and concentration of spectral projectors of sample covariance," *The Annals of Statistics*, vol. 45, no. 1, pp. 121–157, 2017.

[3] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

[4] Z. D. Bai and Y. Q. Yin, "Convergence to the semicircle law," *The Annals of Probability*, pp. 863–875, 1988.

[5] Y. Yin, Z. Bai, and P. Krishnaiah, "Limiting behavior of the eigenvalues of a multivariate f matrix," *Journal of multivariate analysis*, vol. 13, no. 4, pp. 508–516, 1983.

[6] J. W. Silverstein, "Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 55, no. 2, pp. 331–339, 1995.

[7] N. El Karoui *et al.*, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *The Annals of Statistics*, vol. 36, no. 6, pp. 2757–2790, 2008.

[8] Z. Bai, J. Chen, and J. Yao, "On estimation of the population spectral distribution from a high-dimensional sample covariance matrix," *Australian & New Zealand Journal of Statistics*, vol. 52, no. 4, pp. 423–437, 2010.

[9] O. Ledoit, M. Wolf, *et al.*, "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *The Annals of Statistics*, vol. 40, no. 2, pp. 1024–1060, 2012.

[10] O. Ledoit and M. Wolf, "Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions," *Journal of Multivariate Analysis*, vol. 139, pp. 360–384, 2015.

[11] Z. Burda, A Görlich, A. Jarosz, and J. Jurkiewicz, "Signal and noise in correlation matrix," *Physica A: Statistical Mechanics and its Applications*, vol. 343, pp. 295–310, 2004.

[12] J. Cooley, "An improved eigenvalue corrector formula for solving the schrödinger equation for central fields," *Mathematics of Computation*, vol. 15, no. 76, pp. 363–374, 1961.

[13] D. J. Graham and N. G. Midgley, "Graphical representation of particle shape using triangular diagrams: An excel spreadsheet method," *Earth Surface Processes and Landforms*, vol. 25, no. 13, pp. 1473–1477, 2000.

[14] I. T. Jolliffe, "Principal components in regression analysis," in *Principal component analysis*, Springer, 1986, pp. 129–155.

[15] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[16] A. Malhi and R. X. Gao, "Pca-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517–1525, 2004.

[17] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *2010 international conference on system science, engineering design and manufacturing informatization*, IEEE, vol. 1, 2010, pp. 27–30.

[18] Q Guo, W Wu, D. Massart, C Boucon, and S De Jong, "Feature selection in principal component analysis of analytical data," *Chemometrics and Intelligent Laboratory Systems*, vol. 61, no. 1-2, pp. 123–132, 2002.

[19] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.

[20] J. C. Gomez and M.-F. Moens, "Pca document reconstruction for email classification," *Computational Statistics & Data Analysis*, vol. 56, no. 3, pp. 741–751, 2012.

[21] M Zahedi and A. G. Sorkhi, "Improving text classification performance using pca and recall-precision criteria," *Arabian Journal for Science and Engineering*, vol. 38, no. 8, pp. 2095–2102, 2013.

[22] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[23] P. Nedungadi, H. Harikumar, and M. Ramesh, "A high performance hybrid algorithm for text classification," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, IEEE, 2014, pp. 118–123.

[24] W. Li, J. Chen, Y. Qin, Z. Bai, and J. Yao, "Estimation of the population spectral distribution from a large dimensional sample covariance matrix," *Journal of Statistical Planning and Inference*, vol. 143, no. 11, pp. 1887–1897, 2013.

[25] C. Stein, "Estimation of a covariance matrix," in *39th Annual Meeting IMS, Atlanta, GA, 1975*, 1975.

[26] P. J. Bickel, E. Levina, *et al.*, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[27] M. Gavish and D. L. Donoho, "Optimal shrinkage of singular values," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2137–2152, 2017.

[28] W. Kong, G. Valiant, *et al.*, "Spectrum estimation from samples," *The Annals of Statistics*, vol. 45, no. 5, pp. 2218–2247, 2017.

[29] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 24–33, 2014.

[30] S. Günal, S. Ergin, M. B. Gülmezoğlu, and Ö. N. Gerek, "On feature extraction for spam e-mail detection," in *International Workshop on Multimedia Content Representation, Classification and Security*, Springer, 2006, pp. 635–642.

[31] I. Idris, A. Selamat, and S. Omatu, "Hybrid email spam detection model with negative selection algorithm and differential evolution," *Engineering Applications of Artificial Intelligence*, vol. 28, pp. 97–110, 2014.

[32] A. Uysal, S Gunal, S Ergin, and E. S. Gunal, "The impact of feature extraction and selection on sms spam filtering," *Elektronika ir Elektrotechnika*, vol. 19, no. 5, pp. 67–72, 2013.

[33] X. Feng, H. Qin, Q. Shi, Y. Zhang, F. Zhou, H. Wu, S. Ding, Z. Niu, Y. Lu, and P. Shen, "Chrysin attenuates inflammation by regulating m1/m2 status via activating pparγ," *Biochemical pharmacology*, vol. 89, no. 4, pp. 503–514, 2014.

[34] E. Saraç and S. A. Özel, "An ant colony optimization based feature selection for web page classification," *The Scientific World Journal*, vol. 2014, 2014.

[35] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[36] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive bayes model," in *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2013, pp. 194–201.

[37] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive bayes for text categorization," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 9, pp. 2508–2521, 2016.

[38] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert systems with Applications*, vol. 43, pp. 82–92, 2016.

[39] N. Nguyen, K. Yamada, I. Suzuki, and M. Unehara, "Hierarchical scheme for assigning components in multinomial naive bayes text classifier," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, 2018, pp. 335–340.

[40] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[41] P. Langley, W. Iba, K. Thompson, *et al.*, "An analysis of bayesian classifiers," in *Aaai*, vol. 90, 1992, pp. 223–228.

[42] J. Su, J. S. Shirab, and S. Matwin, "Large scale text classification using semi-supervised multinomial naive bayes," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Citeseer, 2011, pp. 97–104.

[43] Venkatesh and K. V. Ranjitha, "Classification and optimization scheme for text data using machine learning naïve bayes classifier," in *2018 IEEE World Symposium on Communication Engineering (WSCE)*, 2018, pp. 33–36.

[44] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[45] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, Springer, 1998, pp. 137–142.

[46] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101*, 2016.

[47] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[48] R. Albright, "Taming text with the svd," *SAS Institute Inc*, 2004.

[49] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[50] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[51] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 26–39, 2016.

[52] L. Zhang, L. Jiang, C. Li, and G. Kong, "Two feature weighting approaches for naive bayes text classifiers," *Knowledge-Based Systems*, vol. 100, pp. 137–144, 2016.

[53] L. Jiang, Z. Cai, H. Zhang, and D. Wang, "Naive bayes text classifiers: A locally weighted learning approach," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 25, no. 2, pp. 273–286, 2013.

[54] J. Chen, H. Matzinger, H. Zhai, and M. Zhou, "Centroid estimation based on symmetric kl divergence for multinomial text classification problem," *arXiv preprint arXiv:1808.10261*, 2018.

[55] Y. Wu, "A new instance-weighting naive bayes text classifiers," in *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, 2018, pp. 198–202.

[56] G. Feng, J. Guo, B.-Y. Jing, and T. Sun, "Feature subset selection using naive bayes for text classification," *Pattern Recognition Letters*, vol. 65, pp. 109–115, 2015.

[57] D. Mladenic and M. Grobelnik, "Word sequences as features in text-learning," in *In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98*, Citeseer, 1998.

[58] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1265–1287, 2003.

[59] J. M. G. Hidalgo and M. d. B. Rodriguez, "Integrating a lexical database and a training collection for text categorization," *arXiv preprint cmp-lg/9709004*, 1997.

[60] F. Li and Y. Yang, "A loss function analysis for classification methods in text categorization," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 472–479.

[61] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*, vol. 1, 1999, pp. 61–67.

[62] D. D. Lewis, *Reuters-21578*.

[63] K. Lang, *20 newsgroups data set*.

[64] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, ACM, 1998, pp. 148–155.

[65] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1998, pp. 90–95.

[66] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.

[67] K.-M. Schneider, "A new feature selection score for multinomial naive bayes text classification based on kl-divergence," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Association for Computational Linguistics, 2004, p. 24.

[68] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *IJCAI*, vol. 3, 2003, pp. 587–592.

[69] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1501–1536, 2011.

[70] N. Nguyen and R. Caruana, "Classification with partial labels," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2008, pp. 551–559.

[71] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Advances in neural information processing systems*, 2003, pp. 921–928.

[72] M.-L. Zhang and F. Yu, "Solving the partial label learning problem: An instance-based approach," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[73] L. Feng and B. An, "Partial label learning with self-guided retraining," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3542–3549.

[74] Q.-W. Wang, Y.-F. Li, and Z.-H. Zhou, "Partial label learning with unlabeled data," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, AAAI Press, 2019, pp. 3755–3761.

[75] N. Xu, J. Lv, and X. Geng, "Partial label learning via label enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5557–5564.

[76] M.-L. Zhang, B.-B. Zhou, and X.-Y. Liu, "Partial label learning via feature-aware disambiguation," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1335–1344.

[77] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, Citeseer, vol. 752, 1998, pp. 41–48.

[78] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2008, pp. 614–622.

[79] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[80] ——, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[81] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[82] A. McCallum, "Multi-label text classification with a mixture model trained by em," in *AAAI workshop on Text Learning*, 1999, pp. 1–7.

[83] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, IBM New York, vol. 3, 2001, pp. 41–46.

[84] Jeff, *Dimension reduction with pca*.

[85] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, 2000.

[86] O. Alter and G. H. Golub, "Singular value decomposition of genome-scale mrna lengths distribution reveals asymmetry in rna gel electrophoresis band broadening,"

Proceedings of the National Academy of Sciences, vol. 103, no. 32, pp. 11 828–11 833, 2006.

[87] N. M. Bertagnolli, J. A. Drake, J. M. Tennessen, and O. Alter, "Svd identifies transcript length distribution functions from dna microarray data and reveals evolutionary forces globally affecting gbm metabolism," *PloS one*, vol. 8, no. 11, e78913, 2013.

[88] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.

[89] L. Omberg, J. R. Meyerson, K. Kobayashi, L. S. Drury, J. F. Diffley, and O. Alter, "Global effects of dna replication and dna replication origin activity on eukaryotic gene expression," *Molecular systems biology*, vol. 5, no. 1, p. 312, 2009.

[90] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the davis–kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2014.

[91] T. W. Anderson, "An introduction to multivariate statistical analysis," Wiley New York, Tech. Rep., 1962.